Neue Methoden zur Entdeckung von Fehlspezifikation bei Latent-Trait-Modellen der Veränderungsmessung

D I S S E R T A T I O N

zur Erlangung des akademischen Grades
doctor rerum naturalium
(dr. rer. nat.)
im Fach Psychologie

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät I
Humboldt-Universität zu Berlin

von
Herr Dipl.-Stat. Stefan Klein
geboren am 12.3.1967 in München

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jürgen Mlynek

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:
Prof. Dr. Elmar Kulke

Gutachter:

1. Prof. Dr. Bodo Krause
2. Prof. Dr. Tenko Raykov
3. Prof. Dr. Karl Christoph Klauer

eingereicht am:	23. November 2002
Tag der mündlichen Prüfung:	9. Mai 2003

Abstract

In this thesis, new methods are developed for the detection of misspecification within Linear Logistic Test Models (=LLTM) and similar model classes for the measurement of change. The phrase "misspecification" will be used if a wrong selection of latent traits is chosen for the estimation of the LLTM. Misspecification can lead to erronious estimation ([Baker, 1993]). Using the newly developed methods, it is possible to measure the extent of deviations between the proposed model and the data. This can be done without using estimated parameter values.
First a method is introduced which is based on the well-known Mantel-Haenszel-test. For some hypotheses, this method can be used instead of a Likelihood Ratio Test (e.g. [Fischer, 1995b]).
The Main topic of this thesis are uniformly most powerful tests for the measurement of person fit and related effect measures. These effect measures can be used for the identification of subpopulations where the proposed model does not hold. Statistical properties of these tests resp. effect measures are examined by simulations and power calculations using the SAS software. Furthermore, examples of the application of these methods are given.

Keywords:
Item Response Theory, LLTM, Person fit test, SAS

Zusammenfassung

Ziel der Arbeit ist die Entwicklung von Modellen zur Entdeckung von Fehlspezifikation im Linear Logistic Test Model (= LLTM) und verwandten Modellen der Veränderungsmessung. Fehlspezifikation bedeutet hierbei, dass dem Modell ein unzutreffendes Muster latenter Traits zugrundegelegt wurde. Dies kann, vgl. z.B. [Baker, 1993], zu bedeutenden Schätzfehlern führen. Die hier vorgestellten Methoden ermöglichen es unter leicht zu erfüllenden Annahmen, Aussagen über das Ausmaß der Unkorrektheit der verwendeten Modellspezifikation zu machen, ohne die in der Modellschätzung bestimmten Parameterwerte verwenden zu müssen. Zunächst wird eine auf dem Mantel-Haenszel-Test beruhende Methodik vorgestellt, die bei Tests bezüglich der Veränderungsparameter eines LLTMs als direkte Konkurrenz zu den bekannten Likelihood-Ratio-Tests für das LLTM anzusehen ist, wie sie z.B. bei [Fischer, 1995b] vorgestellt werden. Weiterhin werden für das LLTM optimierte Personenfittests und daraus abgeleitete Effektgrößen vorgestellt. Diese ermöglichen das Auffinden von Subpopulationen, bei denen eine Abweichung zum angenommenen Modell aufgetreten ist. Es werden die statistischen Eigenschaften dieser Tests resp. Effektgrößen mittels Simulation und Teststärkeberechnung untersucht und Anwendungsbeispiele für diese Methoden vorgestellt.

Schlagwörter:
Item-Response-Theorie, LLTM, Personenfittests, SAS

Danksagung

An dieser Stelle möchte ich mich bei meinem Betreuer, Prof. Dr. Bodo Krause, für viele fruchtbare und wertvolle Diskussionen sowie bei Dr. Stefan Künstner für die Überlassung des EORTC QLQ-C30-Datensatzes bedanken.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

Die vorliegende Arbeit beschäftigt sich mit der Methodik von Veränderungsmessungen. Veränderungsmessungen wurden in der empirischen Psychologie schon sehr früh eingesetzt, z.B. bei Ebbinghaus 1897: Hier wurde an Breslauer Gymnasiasten eine Längsschnittstudie durchgeführt, bei der das Ausmaß sprachlicher und arithmetischer Fähigkeiten an verschiedenen Tageszeiten untersucht wurde. Ähnlich früh begannen Psychologen, die bei der Veränderungsmessung auftretenden Probleme zu untersuchen. Benannt werden in der psychometrischen Literatur z.B. das Anfangswertgesetz (vgl. [Wilder, 1931]), die geringe Reliabilität von Differenzmaßen der Veränderungsmessung (vgl. hierzu auch [Webster and Bereiter, 1963]) sowie die Problematik von Boden- und Deckeneffekten (vgl. [Bereiter, 1963]).

Bei manchen Autoren führte das Vorhandensein dieser Probleme dazu, dass von der Veränderungsmessung explizit abgeraten wurde (vgl. z.B. [Cronbach and Furby, 1970]). Diese Probleme sind vor allem in der klassischen (psychologischen) Testtheorie von Bedeutung (vgl. z.B. [Raykov, 1987], [Raykov, 1995]). Durch Übergang zu Fehler-in-den-Variablenmodellen (vgl. [Raykov, 1987]), LISREL-Modellen (vgl. [Raykov, 1995]) oder Latent-Trait-Modellen der probabilistischen Testtheorie (vgl. hierzu auch [Fischer, 1995a], [Fischer, 1995b], [Fischer, 1995c]) kann das Ausmaß dieser Probleme verringert werden.

Die vorliegende Arbeit beschäftigt sich mit dem Einsatz von Methoden der Latent-Trait-Theorie für die Veränderungsmessung, vor allem mit den von G. H. Fischer vorgestellten Modellen (vgl. z.B. [Fischer, 1995b], [Fischer, 1995a]), die die Messung von Veränderungen in einer Subpopulation ermöglichen. Die oben angesprochenen Problemkomplexe werden durch diese Modelle gelöst bzw. verringert (vgl. [Fischer, 1995c]): Durch Verwendung der logistischen Link-Funktion kann man z.B. die gesamte Zahlengerade als Wertebereich für den Schätzwert eines Veränderungsparameters ausnutzen, wodurch die Problematik von Boden- und Deckeneffekten entfällt. Weiterhin benötigt man keine Differenzwerte als Schätzungen für die Veränderung, wodurch die Problematik der niedrigen Reliabilität von Differenzmessungen entfällt. Erkauft werden diese Vorteile durch den Verzicht auf die Messung personenspezifischer Veränderungen. In diesen Modellen können Veränderungen

nur für Subpopulationen geschätzt werden.

Im Gegensatz zum klassischen Rasch-Modell können die durch Fischer vorgestellten Modelle auch mehrere Traits gleichzeitig modellieren. Entscheidend bei diesen Veränderungsmessmodellen ist daher die richtige Zuordnung zwischen Personen und Veränderungstraits. An die von Fischer vorgestellten Modelle knüpft die vorliegende Arbeit an, indem Methoden vorgestellt werden, mit denen die Gültigkeit der Zuordnung Person/Trait untersucht werden kann.

Zu diesem Zweck wird der Begriff der Fehlspezifikation eingeführt. In der Latent-Trait-Theorie wird der Begriff (vgl. [Baker, 1993]) in folgendem Zusammenhang verwendet: Loglineare Raschmodelle (vgl. [Fischer, 1995b], [Fischer, 1995a]) erlauben die gleichzeitige Verwendung mehrerer unterschiedlicher Traits. Falls für die Messung bei Person v ein Subtrait i angenommen wird, Person v jedoch in Wirklichkeit einen Trait j verwendet, so wurde das zugrunde liegende Modell falsch konzipiert, und es liegt Fehlspezifikation des Modells vor.

Um Fehlspezifikation entdecken zu können, werden Techniken aus der (Rasch-)Testtheorie eines Messzeitpunkts auf Untersuchungsdesigns mit mehreren Zeitpunkten übertragen. Zum einen handelt es sich hier um die Untersuchung auf Item Bias (vgl. [Fischer, 1993]), zum anderen um die Verwendung von Personenfittests (vgl. z.B. [Klauer, 1995]).

Aus der Untersuchung auf Item Bias entwickeln wir eine Methode, mit der festgestellt werden kann, ob ein Veränderungsmesswert bei einem Item in einer bestimmten Subpopulation den gleichen Wert aufweist wie der Veränderungsmesswert dieses Items in anderen Subpopulationen. Die in dieser Arbeit vorgestellten Techniken sind unabhängig von den Itemparameterschätzungen und daher besonders zur Aufdeckung von Fehlspezifikation geeignet.

Die zweite in dieser Arbeit verwendete Methodenklasse ist die Technik der Personenfitindizes, mit denen untersucht wird, wie gut das Antwortmuster einer Person zu einem vorgegebenen Item-Response-Theory-Modell (= IRT-Modell) passt. Diese Methodenklasse ist als Ergänzung zu den bekannten Testverfahren in der IRT zu sehen, wie z.B. dem Likelihood-Ratio-Test. Der Einsatz der Personenfitverfahren kann in ähnlicher Weise erfolgen wie der Einsatz von Residuen in der Regressionsanalyse. Die Abweichung einer Person von dem vorausgesetzten IRT-Modell hilft bei der Entscheidung, ob das Modell für diese Person richtig definiert ist.

Die in dieser Arbeit vorgestellten Verfahren lassen sich somit in zwei Klassen unterteilen: Einerseits ein (auf dem Mantel-Haenszel-Test aufbauendes) Verfahren, das im Bereich der Veränderungsmessung mit dem Linear Logistic Test Model (vgl. [Fischer, 1995b]) in Konkurrenz zu den klassischen Testverfahren (wie z.B. dem Likelihood-Ratio-Test) treten

kann und dabei eine mögliche Fehlerquelle der klassischen Testverfahren vermeidet.

Andererseits stellen wir Verfahren vor, die (durch die Bereitstellung personenbezogener Abweichungen von einer Nullhypothese) eine wichtige Ergänzung zu den klassischen Testverfahren sein können. Als wichtigen Anwendungsbereich dieser zweiten Gruppe von Verfahren sehen wir dabei die Evaluation von Psychotherapien. Unserer Auffassung nach kann mit den in dieser Arbeit entwickelten Verfahren der Einsatz des Linear Logistic Test Model bei der Evaluation von Psychotherapien erleichtern. Zudem können durch den Einsatz unserer Methoden manche Probleme vermieden werden, die bei den gegenwärtigen Evaluationsstandards häufig auftreten. Daher beschreiben wir zunächst in Kapitel 2.1 den derzeitigen Stand bei der Evaluation von Psychotherapien und erläutern in Kapitel 12.5, inwieweit die hier vorgestellten Methoden zur Verbesserung der derzeitigen Situation beitragen können.

Kapitel 2

Modelle und Probleme der Veränderungsmessung

In diesem Kapitel beschäftigen wir uns zunächst mit Problemen bei der Evaluation in der Psychotherapieforschung. Danach stellen wir die verwendeten Modelle der Veränderungsmessung vor, um schließlich das Phänomen „Fehlspezifikation“ näher zu beschreiben, welches das eigentliche Thema dieser Arbeit ist.

2.1 Standards zur Evaluation in der Psychotherapieforschung

Kritik an den Evaluationsstandards in der Psychotherapieforschung

An dieser Stelle fassen wir die von [Metzler and Krause, 1997] und [Hager, 1998] vorgebrachte Kritik an den Evaluationsstandards der Psychotherapieforschung zusammen. Metzler/Krause fordern die Anwendung von Standards der (aus der Pharmaforschung stammenden) „Good Clinical Practice“ (= GCP) auch in Studien zur Psychotherapieforschung. Diese sollen allerdings in leicht abgewandelter Form zum Einsatz kommen: Analog zur Pharma-Forschung wird ein 4-Phasen-Schema vorgeschlagen. Unterschiede zur GCP bestehen kaum und sind lediglich auf die unterschiedlichen Untersuchungsmethoden in Medizin und Psychologie zurückzuführen. Die eigentliche Evaluationsuntersuchung findet nach diesem Vorschlag in der 3. Phase statt, während die ersten beiden Phasen Voruntersuchungen beinhalten, und die 4. Phase erst nach der Zulassung der Therapie beginnt. Weiterhin fordern Metzler/Krause bestimmte Mindeststandards für das Design von Evaluationsstudien.

Gefordert werden für Phase 3-Studien

- ein Kontrollgruppendesign als Versuchsplan,
- eine genaue Kontrolle über die verwendeten Patientenpopulationen,
- ein prospektiver Studienaufbau,
- eine randomisierte Zuteilung der Patienten zu Behandlungs- und Kontrollgruppen,
- eine ansatzweise Verblindung von Therapeuten und Patienten, und
- eine Kontrolle der Compliance der Therapie.

Die im Zusammenhang dieser Arbeit wichtigen Gesichtspunkte des Metzler/Krauseschen Ansatzes liegen aber in der Auswahl geeigneter Zielkriterien und einer konfirmatorischen statistischen Analyse dieser Zielkriterien in Verbindung mit sinnvoller Planung des Stichprobenumfangs.

Metzler/Krause plädieren im Hinblick auf die Auswahl des Zielkriteriums für eine Trennung zwischen den eigentlichen Zielkriterien, aufgrund derer die Wirkungsweise einer Therapie beurteilt werden kann, und Surrogatkriterien (d.h. begleitend erhobene Kriterien, die zwar von einem Therapieerfolg beeinflusst werden, aufgrund derer aber kein Rückschluss auf den Therapieerfolg möglich ist). Nach Ansicht von Metzler/Krause werden aufgrund dieser fehlenden Unterscheidung oftmals zu viele Zielkriterien erhoben, was eine genaue Definition des Erfolgs einer Therapie unmöglich macht. In diesem Zusammenhang kritisieren Metzler/Krause auch die Auswertung solcher multivariater Zielmerkmale mit univariaten Methoden. Dies führt dazu, dass aus der Vielfalt erhobener Zielvariablen nur solche ausgewählt werden, bei denen signifikante Effekte nachgewiesen werden konnten. Dies widerspricht einerseits der Philosophie des Signifikanztests und führt andererseits zu größerer Unklarheit bei der Definition der „Wirkung“ einer Therapie.

Im Vergleich mit der Pharmaforschung zeigt sich: Auch dort ist das Phänomen des Surrogatkriteriums bekannt. Surrogatkriterien tauchen dort vor allem in den frühen Evaluationsphasen auf. In späteren Phasen werden Surrogatendpunkte nur unter Vorbehalten verwendet, so etwa

- wenn der wirkliche Endpunkt einer Evaluation ethisch nicht vertretbar ist (z.B. wegen Tod des Patienten) (nach [Harder, 1994]), oder
- wenn „der Endpunkt in dem vor der Zulassung möglichen Beobachtungszeitraum nicht erfasst werden kann“ ([Harder, 1994]).

Der Zusammenhang zwischen Surrogat und wirklichem Endpunkt sollte

> ...durch epidemiologische Daten abgesichert sein, ebenso der klinische Nutzen [...] bei Veränderungen des Surrogatendpunktes durch die Therapie ([Harder, 1994]).

Weiterhin fordern [Metzler and Krause, 1997] eine genauere Planung des Stichprobenumfangs in Therapieevaluationsstudien, als dies bisher der Fall ist. Sie untersuchen die Stichprobenumfänge in der von [Grawe et al., 1994] durchgeführten Metaanalyse. Bei 50% der untersuchten Studien ist der Gesamtstichprobenumfang kleiner als 50, bei 92.2 % der Studien nicht größer als 100. Bei einer stärkeren Verwendung von multivariaten statistischen Methoden müsste der Stichprobenumfang deutlich gegenüber den hier genannten Zahlen erhöht werden. In diesem Zusammenhang diskutieren [Metzler and Krause, 1997] auch mögliche Einflussvariablen bezüglich des Stichprobenumfangs. Im Allgemeinen plädieren Metzler/Krause jedoch für eine möglichst geringe Zahl von Zielgrößen, wodurch eine genaue Planung der Stichprobengröße erst möglich wird.

In der Pharmaforschung wird von ähnlichen Grundsätzen ausgegangen:

> Die Notwendigkeit einer exakten Hypothesendefinition und der adäquaten Fallzahlschätzung werden in 4.3 der Richtlinien zur ordnungsgemäßen Durchführung von klinischen Studien in der Europäischen Gemeinschaft –Good Clinical Practice– ausdrücklich betont ([Aydemir, 1994]).

[Hager, 1998] diskutiert die Forderungen von [Metzler and Krause, 1997] und begrüßt ihren Vorschlag der Setzung von Standards. Allerdings kritisiert er einige wesentliche von Metzler/Krause vorgebrachte Gedanken:

- Hager begrüßt zwar die Forderung nach Kontrollgruppendesigns, sieht jedoch Schwierigkeiten in der Wahl geeigneter Kontrollgruppen.
- Bei der Frage nach der Randomisierung kritisiert Hager, dass eine Randomisierung im psychologischen Umfeld oft nicht möglich ist. In solchen Fällen sollte

  > ...den wichtigsten potenziellen Störfaktoren bzw. Störhypothesen nachgegangen und gezeigt werden, dass die Anwendungsvoraussetzungen für die letzteren nicht vorgelegen haben ([Hager, 1998], S. 73).

- Ansonsten fordert auch Hager eine weitere Verbreitung der Randomisierung in psychotherapeutischen Studien.
- Hauptkritikpunkt von Hager ist die Frage der angemessenen statistischen Verfahren. Hager zweifelt die Verwendung multivariater Verfahren an, da man mit diesen nicht zielgerichtet diejenigen Hypothesen überprüfen könne, an denen man interessiert ist:

  > ...dass die multivariaten Testkriterien üblicherweise dann statistisch signifikant werden, wenn sich die Werte von mindestens einer der AVn statistisch bedeutsam in zwei oder mehr Versuchsgruppen voneinander unterscheiden. An dieser Art von unspezifischer Information ist man aber in der Regel gar nicht interessiert... ([Hager, 1998], S. 75).

An dieser Stelle muss Hager widersprochen werden: Bei der Verwendung multivariater Verfahren muss anschließend untersucht werden, bei welchen Variablen sich statistisch

signifikante Effekte zeigen lassen. Der Sinn des multivariaten Tests liegt lediglich in der Absicherung eines Gesamtsignifikanzniveaus für alle durchgeführten Tests zusammen, so dass Fehlinterpretationen durch zufällig aufgetretene Effekte vermieden werden können.

Methodische Standards in der Therapieevaluation: eine vergleichende Studie

[Sbandi et al., 1993] führten im Auftrag des österreichischen Kultusministeriums eine Untersuchung über die Evaluationsstandards in der Psychotherapie durch. Hierbei wurde lediglich eine Bestandsaufnahme durchgeführt; wertende Aussagen wurden vermieden.

In einem inhaltlichen Teil fassen [Sbandi et al., 1993] einige zu diesem Zeitpunkt erschienene Beiträge zur Evaluationsforschung zusammen. Folgende Aussagen sind u.E. nach wesentlich:

- Bei der Evaluation von Psychotherapien werden häufig Einzelfall-Prozessstudien als zu bevorzugende Methode genannt, da nur durch solche Studien

  > „mehr Klarheit über das komplexe Gefüge der Wirkmechanismen ... erlangt werden“ kann ([Sbandi et al., 1993], S. 11).

  Gruppenuntersuchungen seien im Gegensatz dazu

  > „nur zum Nachweis der globalen Effektivität von Psychotherapie geeignet, nicht aber für die Herausarbeitung der Effekte unterschiedlicher Therapieformen und Elemente“ ([Grawe, 1987], zitiert nach [Sbandi et al., 1993], S. 11).

  Zudem wird kritisiert, dass gruppenstatistische Analysen „verallgemeinern, Unterschiede verwischen“ ([Wittmann, 1984], zit. nach [Sbandi et al., 1993], S. 11). Einzelfallanalysen werden also deswegen gegenüber gruppenstatistischen Methoden bevorzugt, weil man mit diesen die Versuchsumstände genauer berücksichtigen und Wirkprinzipien besser bestimmen kann.

- Zur genauen Kontrolle des Patientenkollektivs bemerken Sbandi et al.:

  > „Therapeutenvariablen wurden nur selten auf ihren Einfluss untersucht... Patientenvariablen wurden zwar häufiger berücksichtigt, aber in erster Linie hinsichtlich leicht feststellbarer Merkmale wie Alter und Geschlecht, kaum jedoch hinsichtlich schwieriger fassbarer Merkmale wie z.B. sozialer Hintergrund oder Aspekte ihres Verhaltens in der psychotherapeutischen Situation“ ([Sbandi et al., 1993], S. 15).

  Auf S. 18 ihrer Untersuchung kommentieren Sbandi et al. pointierter:

  > „Entscheidende Einflussgrößen wie z.B. die Therapeut-Patient-Beziehung werden praktisch nicht oder nur ansatzweise berücksichtigt.“

Wie hieraus ersichtlich, kritisieren Sbandi et al. also die oftmals fehlende Untersuchung des psychosozialen Hintergrundes. [Metzler and Krause, 1997] hatten Ähnliches im Sinn bei ihrer Forderung nach besserer Kontrolle der Compliance einer Therapie.

- Bei der Evaluation psychotherapeutischer Methoden werden für die Beurteilung der Wirksamkeit oft sehr schwache Kriterien verwendet:

  > „Verbesserung der Leitsymptomatik ... Entwicklung von emotionaler Einsicht, Bewusstmachung der latenten Zusammenhänge der intrapsychischen Konfliktlage, Fähigkeit der Intimität. 'Harte Daten', wie z.B. Medikamentenverbrauch ... wurden selten herangezogen."

  Diese „Kriterien" wurden bei der Evaluation psychoanalytischer Methoden verwendet (vgl. [Sbandi et al., 1993], S. 22). Andere Therapiezweige verwendeten härtere, meist symptombezogene Kriterien. Durch die eigene Theorie definierte Erfolgskriterien existieren z.B. bei unspezifischeren Therapieformen, wie der Psychoanalyse und der klientenzentrierten Psychotherapie, dagegen gar nicht bei einer symptomspezifisch ansetzenden Therapie, wie der Verhaltenstherapie (vgl. [Sbandi et al., 1993], S. 22-36). Insgesamt kommen die Autoren zu folgendem Schluss, was die Frage der Effektivität von Psychotherapien betrifft:

  > „In der ... Literatur der letzten 10 Jahre wird zwar davon ausgegangen, dass es spezifische Wirkungen der Psychotherapie gibt. Unklar ist aber, was genau das Spezifische ist. Im Einzelfall ist kaum zu unterscheiden zwischen Psychotherapie-Variablen und Variablen des sozialen Umfelds, da Psychotherapie immer in einem sozialen Umfeld stattfindet, und sich dieses Umfeld – im Sinne eines klassischen Experiments – nicht stabil halten lässt."

Folgerung

Aus den angeführten Artikeln von Metzler/Krause sowie von Sbandi et al. lässt sich ein deutliches Unbehagen an den derzeitigen Evaluationsstandards in der Psychotherapieforschung herauslesen. Als wichtigster Punkt hierbei stellt sich die Frage der Messung (resp. Messbarkeit?) der Effektivität einer Psychotherapie. In dieser Arbeit werden wir Methoden vorstellen, mit denen genauere Aussagen über die Wirkungen einer Psychotherapie möglich sind, wenn man multivariate Methoden (in diesem Fall: multivariate Rasch-Modelle) verwendet. In diesem Sinn können die in dieser Arbeit vorgestellten Methoden ein Schritt zur besseren Beurteilung der Effektivität von Psychotherapien sein.

2.2 Latent-Trait-Modelle der Veränderungsmessung

In dieser Arbeit geht es um Veränderungsmessung mit Latent-Trait-Modellen, d.h. um Modelle, die die Wahrscheinlichkeit für die richtige Beantwortung eines Items durch eine Versuchsperson mit Hilfe einer latenten (Fähigkeits-)Variable, dem sog. Trait bestimmen.

Die in dieser Arbeit verwendeten Modelle sind Abarten des Rasch-Modells ([Rasch, 1960])

$$P\left(X_{vi}=1\right)=\frac{\exp\left(\theta_v-\beta_i\right)}{1+\exp\left(\theta_v-\beta_i\right)} \tag{2.1}$$

bei dem sich die Lösungswahrscheinlichkeit $P\left(X_{vi}=1\right)$ eines Items i alleine durch den latenten Fähigkeitsparameter θ_v der Versuchsperson v und den Schwierigkeitsparameter β_i des Items berechnen lässt. Fähigkeits- und Schwierigkeitsparameter sind additiv verknüpft, falls man den Logit der Lösungswahrscheinlichkeit betrachtet. Als Erweiterungen des Rasch-Modells für die Veränderungsmessung werden das Linear Logistic Test Model (= LLTM, [Fischer, 1972] bzw. [Fischer, 1995b]), das Linear Logistic model with Relaxed Assumptions (= LLRA, [Fischer, 1995a]) und die Verallgemeinerung dieses Modells auf polytome Items, das sog. LPCM (= Linear Partial Credit Model) betrachtet.

Das Linear Logistic Test Model

Das LLTM (= Linear Logistic Test Model) ist eine multivariate Erweiterung des dichotomen Rasch-Modells. Dabei wird der Schwieigkeitsparameter β_i des Rasch-Modells mit mehreren Subpopulations- oder Subtrait-Parametern α_l verknüpft, so dass

$$\beta_i=\sum w_{il}\alpha_l+c.$$

Hierbei sind die Zahlen w_{il} Elemente einer Gewichtsmatrix $\mathbf{W}$, durch die der p-dimensionale Vektor β der Itemschwierigkeiten mit dem k-dimensionalen Vektor α der Traitparameter verknüpft wird:

$$\beta=\mathbf{W}\alpha+c\mathbf{1}_k\ .$$

c ist eine Konstante, $\mathbf{1}_k$ ein k-dimensionaler Spaltenvektor, der in allen Dimensionen den Wert 1 annimmt. Um eine eindeutige Schätzbarkeit der Parameter zu gewährleisten, muss die Matrix $(\mathbf{W};\mathbf{1}_k)$ den vollen Spaltenrang $p+1$ besitzen (für diese Zusammenfassung: vgl. [Fischer, 1995b], [Bechger et al., 2002]).

Ein LLTM für die Veränderungsmessung erhält man so z.B. durch

$$P(X_{vit}=1)=\frac{exp\left[b_{tgi}(\theta_v-\beta_i+\sum\delta_{gt})\right]}{1+exp\left[b_{tgi}(\theta_v-\beta_i+\sum\delta_{gt})\right]}\ . \tag{2.2}$$

Dabei ist δ_{gt} die Veränderung der Itemschwierigkeit für Subtrait (bzw. Subpopulation) g und Zeitpunkt t. β_i kann dabei als die Itemschwierigkeit im Zeitpunkt 1 (bzw. allgemeiner: in einer Referenzgruppe) interpretiert werden. Für dieses Modell ist Conditional

Maximum Likelihood- (= CML-) Schätzung möglich, da der Summenscore $x_{v\cdot}$ einer Person v suffizient für den Fähigkeitsparameter ist, und somit aus der Likelihood herausgerechnet werden kann (vgl. auch [Fischer, 1995b]). Es kann eine gruppenspezifische Veränderung durch den Parameter δ_{gt} geschätzt werden, nicht jedoch eine personenspezifische Veränderung. X_{vit} steht für die dichotome Zufallsvariable „Ergebnis der Person v bei Item i im Zeitpunkt t“. Die Notation aus (2.2) wird auch im Rest der Arbeit verwendet.

Das LLTM ist genügend flexibel, um zu gewährleisten, dass nicht zu allen Zeitpunkten und in allen Gruppen die gleichen Items dargereicht werden müssen. Dies wird durch die Verwendung der Indikatorvariablen b_{tgi} ermöglicht, die den Wert 1 annimmt, wenn in einer Subpopulation g am Zeitpunkt t das Item i dargeboten wurde, und den Wert 0 andernfalls. Bis auf gewisse Link-Items, die die Verbindung zu den vorhergehenden Zeitpunkten sicherstellen, können somit zu jedem Zeitpunkt andere Items verwendet werden.

Als Kritikpunkt am Einsatz dieses Modells bei der Messung von Veränderungen wurde von [Fischer, 1995a] angemerkt, dass das Modell nicht spezifisch für eine Messung von Veränderungen eingerichtet ist; die Schwierigkeitsparameter der Items sind als Nuisance-Parameter anzusehen. Weiterhin kann ein mangelnder Fit des LLTMs an die Daten mit der Postulierung eines additiven Zusammenhangs zwischen Itemschwierigkeit und Fähigkeit im LLTM zusammenhängen. Durch das Vorhandensein einer solchen, u.U. unnötigen Modellrestriktion wird die Breite des Einsatzbereichs des LLTMs in der Veränderungsmessung verringert. [Fischer, 1995a] postuliert daher ein anderes Modell für die Veränderungsmessung, das sogenannte LLRA (= Linear Logistic Model with Relaxed Assumptions).

Das Linear Logistic model with Relaxed Assumpions

Das Linear Logistic Model with Relaxed Assumptions (= LLRA) unterscheidet sich von dem gerade vorgestellten LLTM durch das Fehlen einer additiven Beziehung $\theta_v - \beta_i$ zwischen Fähigkeits- und Schwierigkeitsparameter. Dadurch bleibt einerseits als (bei CML-Schätzung) einziger zu schätzender Parameter die Veränderung δ_{gt} in Gruppe g zum Zeitpunkt t übrig, andererseits ist das Modell auch dann verwendbar, wenn die Daten eigentlich nicht zu einem Rasch-Modell passen würden. Diese breite Anwendbarkeit wird mit dem Nachteil erkauft, dass man keine Information über Itemschwierigkeiten aus dem Modell erhält.

Das LLRA entsteht aus dem LLTM, wenn man den Ausdruck $\theta_v - \beta_i$ in Gleichung (2.2) durch den Ausdruck θ_{vi} ersetzt, der die Fähigkeit einer Person v bei Item i beschreibt. Somit wird in diesem Modell die Eindimensionalität des Rasch-Modells aufgegeben.

Grundgleichung des Modells ist

$$P(X_{vit} = x_{vit}) = \frac{\exp\left[x_{vit}(\theta_{vi} - \delta_{gt})\right]}{1 + exp\left[(\theta_{vi} - \delta_{gt})\right]} \,. \tag{2.3}$$

Wie man sieht, entspricht diese Gleichung der Grundgleichung eines normalen Rasch-Modells, wenn man als Fähigkeitsparameter einer „virtuellen“ Person vi den Wert θ_{vi} verwendet, sowie als Schwierigkeitsparameter eines „virtuellen“ Items gt den Veränderungsparameter δ_{gt}. „Virtuell“ bedeutet hier, dass die Paare „Item/Person“ als Konstrukte anzusehen sind.

[Fischer, 1995a] empfiehlt die Anwendung des LLRAs mit der Begründung, dass durch dieses Modell (verglichen mit dem LLTM) die messtheoretische Forderung von [Rasch, 1960] nach einer Konzentration auf den zu messenden Parameter erfüllt wird. Andererseits wird dadurch eine Analyse von Itemschwierigkeiten unmöglich. Eine gemeinsame Analyse von Itemschwierigkeiten und Veränderungen ist daher mit dem LLRA nicht möglich. Zudem müssen an jedem Messzeitpunkt die gleichen Items verwendet werden, die schon bei den vorhergehenden Zeitpunkten verwendet worden waren. Dadurch kann es bei Personen, die sich an die Lösung mancher Items erinnern, zu Verzerrungen kommen. Vom messtheoretischen Standpunkt zu bevorzugen wäre ein Verfahren, das zumindestens die Möglichkeit offen lässt, manche Items nur bei bestimmten Zeitpunkten darzureichen. Zusammengenommen bewirken diese beiden Einschränkungen, dass Zusammenhänge zwischen Itemschwierigkeit und gemessener Veränderung im LLRA nicht berücksichtigt werden können.

Das Linear Partial Credit Model (= LPCM)

Das Linear Partial Credit Model (= LPCM) ist eine Verallgemeinerung des LLTMs auf polytome Items, vgl. [Fischer and Ponocny, 1994], [Andersen, 1995]. Es kann aus dem Partial Credit Model ([Masters, 1982], [Andersen, 1995])

$$P\left(X_{vi}=x_{vi}\right)=\frac{\exp\left(x_{vi}\theta_v+\sum_{h\leq x_{vi}}\tau_{ih}\right)}{\sum\exp\left(x_{vi}\theta_v+\sum_{h\leq x_{vi}}\tau_{ih}\right)} \tag{2.4}$$

hergeleitet werden, indem man im Partial Credit-Modell die Schwellenparameter τ_{ih} durch Linearkombinationen zugrunde liegender Parameter α_{lh} ersetzt:

$$\tau_{ih}=w_{i0}+\sum_{l=1}^{p}w_{lh}\alpha_{lh}\ . \tag{2.5}$$

p ist dabei die Gesamtzahl der zugrunde liegenden Basisparameter α_{lh}, die w_{ih} sind (bekannte) Gewichte.

Weitere Klassen von Latent-Trait-Modellen zur Veränderungsmessung

Ein allgemeines Latent-Trait-Modell zur Veränderungsmessung wird von [Meiser, 1996] beschrieben. Dieses Modell kann als Verallgemeinerung von LLTM und LLRM angesehen werden. Zusätzlich sind jedoch auch die Modellierung des Antwortverhaltens polytomer Items sowie die Modellierung itemspezifischer Veränderungen möglich. Schätzungen

können auch in diesem allgemeinen Modell mit Hilfe der Conditional-Maximum-Likelihood-Methode (= CML-Methode) durchgeführt werden. [Meiser, 1996] führt zudem aus, dass dieses allgemeine Veränderungsmodell mit der allgemeinen Theorie loglinearer Modelle verknüpft ist. Daher kann die Schätzung der Modellparameter mit Hilfe normaler statistischer Programmpakete wie z.B. SAS erfolgen.

[Andersen, 1985], [Zwinderman, 1991] bzw. [Hoijtink, 1995] stellen Modelle vor, bei denen die Veränderung mit Hilfe der Fähigkeitsparameter modelliert wird. Dabei wird die Veränderung geschätzt, indem die Fähigkeitsparameter als zufällig aufgefasst werden. Die Verteilung der Fähigkeitsparameter kann an allen Zeitpunkten geschätzt werden. Schätzungen für die Veränderung ergeben sich als Mittelwertdifferenzen zwischen diesen Verteilungen. Kompliziertere Formen dieser Modelle berücksichtigen auch Korrelationen zwischen den einzelnen Zeitpunkten (vgl. [Andersen, 1995]).

Modelle, mit denen man personenspezifische Veränderungen messen kann, werden in [Embretson, 1991] sowie bei [Fischer, 1995c] vorgestellt. Letzterer bezeichnet personenspezifische Veränderungen auch als Modifiability. Embretsons Modell leitet sich, vgl. dazu [Fischer, 1995c], aus einem Rasch-Modell mit unvollständiger Besetzung her. D.h.: Es wird für jede Person v an jedem Zeitpunkt t ein eigener Fähigkeitsparameter θ_{vt} angenommen. An einem Zeitpunkt t ergibt sich die Lösungswahrscheinlichkeit eines Items i daher durch

$$P(X_{vit} = x_{vit}) = \frac{exp[x_{vit}(\theta_{vt} - \beta_i)]}{1 + exp(\theta_{vt} - \beta_i)} \quad . \tag{2.6}$$

Indem man auf die Summenscores der Personen an den einzelnen Zeitpunkten bedingt, kann man mittels CML-Methode die Itemschwierigkeiten schätzen. Die Schätzung der personenspezifischen Veränderungen erfolgt über die Schätzung der Personenparameter bei Zeitpunkt $t = 1$, und dann schrittweise über die Schätzung der Veränderung zum jeweils vorhergehenden Zeitpunkt.

Bei Fischers Modell wird als Grundannahme eine Abart des LLTMs angenommen, und für dieses Modell werden die Schwierigkeitsparameter geschätzt. Die geschätzten Itemparameter sind daher dieselben wie im LLTM. Bei bekannten Schwierigkeitsparametern können die personenspezifischen Veränderungsparameter geschätzt werden. Dies führt zu derselben Likelihood wie in Kapitel 8.1.

Auswahl der verwendeten Modellklasse für die Veränderungsmessung

Im weiteren Verlauf der Arbeit werden wir i.d.R. das LLTM voraussetzen. Für Verallgemeinerungen der in dieser Arbeit beschriebenen Tests werden auch das LLRA sowie das Veränderungsmodell von [Meiser, 1996] berücksichtigt. Bei der Verallgemeinerung der Personenfittests auf polytome Daten fiel die Wahl auf das Modell von [Meiser, 1996], weil

sowohl LLTM, als auch das LLRA, als auch das LPCM Spezialfälle des Meiserschen Modells sind.

In den weiteren Kapiteln werden wir folgende Notation verwenden:

- X_{vit} ist die zu der (von Person v im Zeitpunkt t) gegebenen Antwort auf ein Item i gehörende Zufallsgröße.
- Die Gesamtpopulation besteht aus mehreren Teilpopulationen $g = 1, \ldots, G$. Der Veränderungsparameter zum Zeitpunkt t ($=\delta_{gt}$) ist abhängig von der Zugehörigkeit einer Person zu einer bestimmten Subpopulation. Innerhalb einer Subpopulation g und bei gegebenem Zeitpunkt t ist δ_{gt} i.d.R. konstant, d.h. wir gehen meist von einer für alle Items konstanten Veränderung aus. Falls dies nicht der Fall ist, so wird extra darauf hingewiesen.
- Mit θ_v bezeichnen wir die Fähigkeit einer Person, mit β_i die Schwierigkeit eines Items.

2.3 Fehlspezifikation und Differential Item Functioning

Fehlspezifikation: Die Untersuchung von Modellen auf ihre Gültigkeit

In dieser Arbeit werden neue Methoden zum Erkennen von Fehlspezifikation vorgestellt. Der Begriff „Fehlspezifikation“ stammt aus dem Bereich der „Generalized Linear Models“ (= GLMs) (vgl. auch [Fahrmeir and Tutz, 1994], Kap. 4), die als eine Verallgemeinerung der Klasse der loglinearen Rasch-Modelle (vgl. z.B. [Meiser, 1996]) verstanden werden können.

Der Begriff der Fehlspezifikation wird im Zusammenhang der GLMs mit folgendem Phänomen verbunden: Bei der Anpassung eines Modells aus der Klasse der GLMs kann es unterschiedliche Modellspezifikationen geben, die zu einem annehmbaren Fit des Modells führen. Falls man nur den Modellfit betrachtet, kann man nicht mit Sicherheit entscheiden, welche dieser unterschiedlichen Modellspezifikationen dem „wahren“ Modell entspricht. Falls in der Wirklichkeit nicht zutreffende Modelle verwendet werden, liegt Fehlspezifikation vor. Im Zusammenhang der GLMs kann Fehlspezifikation unterschiedliche Formen annehmen: Es kann eine falsche Link-Funktion verwendet werden, es können falsche Prädiktoren verwendet werden, oder es werden falsche Verteilungsannahmen getroffen.

Bei Voraussetzung eines LLTMs können Fehlspezifikationen nur durch eine falsche Prädiktorenauswahl verursacht werden. Dies ist auch der Zusammenhang, in dem [Baker, 1993] den Begriff Fehlspezifikation verwendet.

Ziel dieser Arbeit ist es, ergänzende Methoden zu den gebräuchlichen Anpassungstests für das LLTM (vgl. z.B. [Fischer, 1995b] oder [Bechger et al., 2002]) vorzustellen, die die Unterscheidung zwischen konkurrierenden Modellen erleichtern. Von besonderem psychologischen Interesse sind unserer Ansicht nach zwei verschiedene Arten von Fehlspezifikation:

1. Gibt es nicht berücksichtigte Subpopulationen, die einen Einfluss auf das Antwortverhalten besitzen?
2. Verhalten sich einzelne Items unter bestimmten vorgegebenen Bedingungen bzw. in bestimmten Subpopulationen anders, als man es aus dem Modell heraus erwarten würde?

Fragestellung 1 wird bei psychologischen Tests auch als „mangelnder Personenfit“ beschrieben. Fragestellung 2 ist die Frage nach der Existenz von Interaktionen zwischen Personen- (bzw. Populations-) Variablen und Itemvariablen. Bei psychologischen Tests mit nur einem Messzeitpunkt wird dieser Sachverhalt auch als „Differential Item Functioning“ oder „Item Bias“ bezeichnet. Interaktionen zwischen zwischen Item und Veränderungsparameter werden wir, da dies ein ähnlich strukturiertes Problem wie das vorhergehende darstellt (und um eine griffigere Bezeichnung zu verwenden), als „Bias bezüglich der Veränderung“ bezeichnen.

Daher werden in dieser Arbeit zwei unterschiedliche Methodenklassen vorgeschlagen: In Kapitel 3 leiten wir aus der Mantel-Haenszel-Technik zur Entdeckung von Item-Bias (vgl. z.B. [Meredith and Millsap, 1992]) Methoden für die Entdeckung von Bias bezüglich der Veränderung ab. Die restlichen Kapitel verfolgen einen zweiten Ansatz: Mit Hilfe von Personenfitindizes wird eine Methodologie zur Entdeckung von Subpopulationen (oder: einzelnen Personen) entwickelt, bei denen die angenommene Traitspezifikation nicht zielführend ist.

Auswirkungen von Fehlspezifikation

Fehlspezifikation bei LLTMs wird mittels einer Simulationsstudie bei [Baker, 1993] untersucht. In dieser Studie werden zu vorgegebenen Parameterwerten zufällig Antwortmuster erzeugt. In unterschiedlichem Ausmaß werden in die vorgegebene Spezifikationsmatrix W Fehler eingebaut. Anschließend werden bei den simulierten Datensätzen die Parameter neu geschätzt. [Baker, 1993] unterscheidet dabei ein LLTM mit einer dicht besetzten Spezifikationsmatrix W sowie ein LLTM mit einer dünn besetzten Spezifikationsmatrix W. Dabei kommt Baker zu dem Ergebnis, dass schon ein leichter Grad an Fehlspezifikation schwerwiegende Auswirkungen auf die Schätzgenauigkeit hat, und zwar sowohl bei dünn als auch bei dicht besetztem W. Die Verwendung einer dicht besetzten Spezifikationsmatrix sowie die Vergrößerung der Schätzstichprobe führten zu besseren, gleichwohl immer noch nicht zufriedenstellenden Resultaten.

Bei [Klein, 1999] wurden Untersuchungen zur Fehlspezifikation für Veränderungsmodelle bei polytomen Items durchgeführt. Dabei wurden in einer Simulation 200 Datensätze zufällig erzeugt, die aus je 100 Probanden bestanden, aufgeteilt auf Versuchs- und Kontrollgruppe. Die Fähigkeitswerte der Probanden entstammten in jedem dieser Datensätze einer Gleichverteilung auf $[-5.6, 2.4]$. Die Schwellenparameter der (dreistufigen) Items lagen im Bereich $[2; 4]$ für die Stufe 2 der Items, sowie im Bereich $[3; 7]$ für die Stufe 3. Diese grundlegenden Schwierigkeiten galten für beide Untersuchungsgruppen. Die Fehlspezifikation wurde mit Hilfe des Veränderungsparameters modelliert. Für jede Person wurde ein eigener Veränderungsparameter modelliert. Die (personenspezifische) Veränderung ergab sich in der Versuchsgruppe durch die Gleichung

$$\delta_{2g} = 2.1 - 0.1 * \theta_v + \epsilon_v \quad , \tag{2.7}$$

mit ϵ_v als $N(0; 0.2)$-verteilter Fehlervariable. In der Kontrollgruppe errechnete sich der (wahre) Veränderungswert durch

$$\delta_{2g} = 0.3 - 0.1 * \theta_v + \epsilon_v \quad . \tag{2.8}$$

Da in den vorgegebenen Untersuchungsgruppen keine konstante Veränderung vorlag, wurde hier Fehlspezifikation im obigen Sinne modelliert.

In der Untersuchung wurden zwei verschiedene Schätzmethoden miteinander verglichen: die nichtparametrische Marginal-Maximum-Likelihood-Methode (= nMML) und die Conditional-Maximum-Likelihood-Methode (= CML). Zu diesen Schätzmethoden vgl. auch [Klein, 1999]. Zunächst wurden dabei die Differenzen der geschätzten Schwellenparameter zu ihren wahren Werten untersucht. Bei fünf von neun zu schätzenden Parametern ergaben sich mit der CML-Schätzung geringere Differenzen zu den wahren Parameterwerten als mit der nMML-Schätzung. Dabei lagen die mittleren Verzerrungen der CML-Schätzung zwischen 0.01 und 1.74, während die mittleren Verzerrungen der nMML-Schätzungen zwischen -2.13 und 3.14 lagen. Für die Schwellenparameter zwischen erster und zweiter Stufe lagen die mittleren Verzerrungen mit der CML-Methode zwischen 0.01 und 0.04, während die mittleren Verzerrungen mit der nMML-Methode zwischen -0.16 und 3.24 schwankten. Die Schätzungen für die Schwellen zwischen zweiter und dritter Kategorie ergaben mit beiden Methoden Verzerrungen von etwa gleicher Größe. Die Standardabweichungen der Verzerrungen bei den Schwellenparametern lagen bei der CML-Methode zwischen 0.47 und 0.85, mit der nMML-Methode zwischen 0.85 und 1.25. Somit kommt die CML-Methode bei der Schätzung der Schwellenparameter zu robusteren Ergebnissen als die nMML-Methode.

Für die Veränderungsparameter konnte keine derartig einfache Untersuchungsmethode verwendet werden, da die wahren Veränderungsparameter ja personenspezifisch waren. Nach den oben angegebenen Formeln lag der Erwartungswert der Veränderung in der Versuchsgruppe bei ca. 2.4, in der Kontrollgruppe bei ca. 0.6. Auch hier kam es zu unbe-

friedigenden Ergebnissen: Der Mittelwert der Schätzungen für die Versuchsgruppe lag bei Verwendung der CML-Methode bei ca. 1.8, während die nMML-Methode einen mittleren Schätzwert von ca. 2.7 lieferte. In der Kontrollgruppe ergab sich mit der CML-Methode eine mittlere Schätzung von ca. 1.1, bei der nMML-Schätzung immerhin noch bei ca. 0.9. Somit lieferte in dieser Simulation bei der Schätzung der Veränderungsparameter die nMML-Methode die besseren Ergebnisse. Der Abstand zwischen Versuchs- und Kontrollgruppe wurde mit der CML-Methode auf im Mittel 0.7 geschätzt, mit der nMML-Methode im Mittel auf 1.8. Wie man sieht, unterschätzte die CML-Methode diesen Abstand deutlich, während die nMML-Methode einen akzeptablen Wert liefert. Zudem wurde mit der CML-Methode in mehreren Fällen in der Kontrollgruppe ein höherer Parameterwert geschätzt als in der Versuchsgruppe.

Festzuhalten bleibt daher, dass beide Schätzverfahren unbefriedigende Resultate ergaben. Fehlspezifikationen in der hier eingeführten massiven Form können bei der Verwendung von LLTM und LLRA in der Veränderungsmessung zu schwerwiegenden Beeinträchtigungen der Schätzgenauigkeit führen. In dieser Arbeit werden daher Methoden entwickelt, mit denen man Personen(gruppen) aufspüren kann, deren Veränderung nicht in das vorausgesetzte Schema passt.

Kapitel 3

Methoden für die Entdeckung von Abweichungen bei einzelnen Items

In diesem Kapitel wird der Themenkreis Item Bias bzw. Bias bezüglich der Veränderung behandelt. Zunächst werden die wichtigsten Methoden zur Erkennung von Item Bias oder auch Differential Item Functioning (= DIF) vorgestellt. Dies sind Methoden zum Erkennen von Item Bias bei Vorliegen nur eines Messzeitpunkts. Anschließend wird anhand eines Artikels von [Roznowski and Reith, 1999] die Bedeutung der Analyse von Item Bias für einen einzigen Zeitpunkt diskutiert. Schließlich wird die Methodik des Mantel-Haenszel-Tests zur Erkennung von Item Bias auf das ähnliche Problem „Bias bezüglich der Veränderung“ übertragen.

3.1 Einleitung

Definition von Differential Item Functioning

Differential Item Functioning (= DIF) oder auch Item Bias wird definiert als das Phänomen, dass aus unterschiedlichen Subpopulationen stammende Personen mit gleichen Fähigkeitswerten unterschiedliche Lösungswahrscheinlichkeiten bei einem gegebenen Item besitzen (vgl. [Lord, 1980]; [Maranon et al., 1997]). Die Wahrscheinlichkeit, ein bestimmtes Item zu lösen, hängt mit verschiedenen äußeren Variablen (z.B. sozioökonomischer Status etc.) zusammen.

Es wird zwischen „Uniform DIF“ und „Non-Uniform DIF“ unterschieden: Uniform DIF tritt immer dann auf, wenn die Lösungswahrscheinlichkeit eines Items i sich für alle Personen einer Subpopulation und für alle Fähigkeitslevels gleichmäßig von der Lösungswahrscheinlichkeit außerhalb der Subpopulation unterscheidet. Uniform DIF lässt sich durch die Lage der Itemantwortfunktionen beschreiben: Die Itemantwortfunktionen bestimmter Subpopulationen sind parallel zueinander, und unterscheiden sich voneinan-

der nur durch die mittlere Position der ICC (= Item Characteristic Curve) auf dem Trait ([Maranon et al., 1997]; [Mellenbergh, 1982]). Non-Uniform DIF ist durch nicht-parallele Itemantwortfunktionen charakterisiert, d.h. Trennschärfe und Schwierigkeit eines Items sind abhängig von der Zugehörigkeit zu bestimmten Subpopulationen (vgl. [Maranon et al., 1997], oder auch [Swaminathan and Rogers, 1990]).

Parallel zum Begriff DIF werden in der Literatur die Begriffe „Measurement Bias“ und „Item Bias“ verwendet. Falls kein DIF vorliegt, spricht man von „Measurement Invariance“ (vgl. auch [Meredith and Millsap, 1992]). Von DIF unterschieden wird der sogenannte „Test Bias“, der als eine Konfundierung zwischen Testergebnis und Zugehörigkeit zu einer bestimmten Subpopulation definiert ist ([Roznowski and Reith, 1999]). Wie diese Autoren betonen, führt DIF nicht zwangsweise zu Test-Bias und somit zu verzerrten Testergebnissen.

Bevor wir näher auf diese Problematik eingehen, beschreiben wir die Mantel-Haenszel-Prozedur zur Entdeckung von DIF.

Die Mantel-Haenszel-Prozedur zur Entdeckung von DIF

Die Mantel-Haenszel-Statistik ist ein Werkzeug zur Entdeckung von DIF unter Gültigkeit des Rasch-Modells. Dabei wird folgendermaßen vorgegangen: Sei Y das Merkmal „Person gehört zu einer Subpopulation g“ und X_i das Merkmal „Lösung des Items i“. Aus den beiden dichotomen Merkmalen kann man nun eine 2×2 –Kontingenztafel bilden. Um das unterschiedliche Verhalten der Personen mit verschiedenen Fähigkeiten zu berücksichtigen, bildet man für jeden aufgetretenen Summenscore eine solche Kontingenztafel. Mit Hilfe dieser Kontingenztafeln wird dann untersucht, ob die Chance, das Item zu lösen, innerhalb der Subpopulation genauso groß ist wie außerhalb der Subpopulation. Unter dem Begriff Chance ist hier der Wettquotient

$$\frac{P(X_i = 1 \mid Y = y)}{P(X_i = 0 | Y = y)} \tag{3.1}$$

zu verstehen. Nullhypothese der Mantel-Haenszel-Prozedur ist, dass der Wettquotient innerhalb der untersuchten Subpopulation in jeder Scoregruppe genauso groß ist wie in der Restpopulation.

Dieses Vorgehen kann gerechtfertigt werden, wenn man das Rasch-Modell mit CML-Schätzung verwendet bzw. ein anderes Modell, dessen Likelihood einer Exponentialfamilie zugehört (vgl. [Fischer, 1993], [Meredith and Millsap, 1992], [Holland and Thayer, 1988]). Somit sollte, wie Fischer betont, dieser Test nur dann angewendet werden, wenn gleiche Trennschärfen für beide Populationen angenommen werden können, also im Fall von Uniform DIF.

3.1.1 Die Zulässigkeit des Mantel-Haenszel-Ansatzes zur Entdeckung von DIF

In diesem Abschnitt geht es um die Frage, wann die Verwendung des Mantel-Haenszel-Ansatzes auf DIF überhaupt zulässig ist. [Meredith and Millsap, 1992] stellen mehrere Bedingungen vor, bei deren Zutreffen die Mantel-Haenszel-Prozedur zulässig ist. Daher sollen die wichtigsten Ergebnisse dieser Arbeit hier kurz zusammengefasst werden. Dabei gehen wir von einem Vektor beobachtbarer Variablen $X = (X_i)$ $(i = 1, ..., m)$ aus. Ein Selektionsmerkmal, das die Gesamtpopulation in zwei oder mehrere Teilmengen aufspaltet, werden wir mit V bezeichnen. Mit Z bezeichnen wir eine Statistik für θ. Weiterhin nehmen wir an, dass die beobachteten Variablen X_i von einer latenten Variable θ abhängen, sowie dass θ, V, X_i nicht stochastisch unabhängig voneinander sind.

Measurement Invariance wird so definiert, dass

> die bedingte Verteilung von X gegeben $\theta = \theta_v$ gleich ist in allen Subpopulationen, die durch verschiedene Realisationen v von V erzeugt werden können, und zwar für alle θ_v mit $Pr(\theta = \theta_v) > 0$

(zitiert und übersetzt nach [Meredith and Millsap, 1992], S. 290).

Meredith/Millsap gelangen dabei zu folgenden Ergebnissen:

- **Measurement Invariance von X gegeben $\theta = \theta_v$ in Bezug auf V** gilt genau dann, wenn X lokal stochastisch unabhängig ist, unter der Bedingung eines vorgegebenen Wertes θ_v für θ (vgl. [Meredith and Millsap, 1992], S. 291ff). Dieses Ergebnis weist zwei Einschränkungen auf: Erstens kann man θ nicht beobachten, zweitens folgt aus Invarianz gegeben $Z = z$ nicht automatisch Measurement Invariance gegeben $\theta = \theta_v$. Um mit der Mantel-Haenszel-Methode DIF entdecken zu können, müssen diese beiden Formen von Invarianz gemeinsam auftreten.
- Eine hinreichende (jedoch nicht notwendige) Bedingung für das gemeinsame Auftreten beider Formen von Invarianz ist die sogenannte **Bayes-Suffizienz** von Z für X, also

$$Pr(X = x, Z = z \mid \theta = \theta_v) = Pr(X = x \mid Z = z)Pr(Z = z \mid \theta = \theta_v) \tag{3.2}$$

  (vgl. auch [Meredith and Millsap, 1992], Theorem 2, Korollare 2.1. bis 2.3.). Dieser Spezialfall ist besonders wichtig, da sich aus ihm einige Implikationen für Latent-Trait-Modelle ableiten.
- Bayes-Suffizienz ist u.a. dann gegeben, wenn $Pr(X_i \mid \theta = \theta_v)$ zur Exponentialfamilie der Verteilungen in $A_i(X_i)$ mit natürlichem Parameter θ gehört, und man

$$Z = \sum_q A_q(X_q) \tag{3.3}$$

setzen kann. Dabei ist Z der Summenscore der $A_i(X_i)$. Daraus folgt die benötigte Bayes-Suffizienz und somit die Gleichheit beider Invarianzformen. Meredith/Millsap beweisen diese Beziehung für den Fall eines eindimensionalen Traits, Generalisierungen auf mehrere Dimensionen sind leicht möglich. Die dargestellten Beziehungen gelten bei der Anwendung des (eindimensionalen) Rasch-Modells, des LLTMs und des LLRAs, nicht jedoch bei der Anwendung des Birnbaum-Modells (vgl. z.B. [Fischer, 1995c], [Meredith and Millsap, 1992]).

3.1.2 Zusammenhang zwischen Item Bias und Testgüte

In diesem Abschnitt soll der Zusammenhang zwischen Item Bias und Testgütekriterien diskutiert werden. Überraschenderweise führt Item Bias oft zu einer Verbesserung der klassischen Testgütekriterien. Dies wird für Messverfahren mit nur einem Zeitpunkt dargestellt.

Ergebnisse einer empirischen Untersuchung zum Zusammenhang zwischen Testgütemaßen und Item-Bias

Item Bias muss nicht immer auch zu schlechteren Testkennwerten führen. Es zeigt sich, dass Tests trotz der Verwendung von durch DIF verzerrten Items immer noch gute Messeigenschaften besitzen können (vgl. [Roznowski and Reith, 1999], bzw. [Roznowski, 1987]). Wenn die Nichttrait-Varianz nur bei wenigen Items einen hohen Einfluss (verglichen mit der Trait-Varianz) besitzt, müssen sich im Summenscore die einzelnen systematischen Fehler der Items nicht unbedingt zu einem Gesamt-Bias addieren, sondern können sich gegenseitig „aufheben“. Falls der Summenscore durch keine bestimmte Ursache des Bias dominiert wird, und somit viele unterschiedliche Subpopulationen betroffen sind, sehen [Roznowski and Reith, 1999] einen Item Bias nicht als nachteilig an.

Dieses Verhalten zeigt sich bei einer Untersuchung, die [Roznowski and Reith, 1999] an ca. 2100 High School-Schülern durchführten, wobei als Vergleichskriterium für die Untersuchungen u.a. der SAT (= Scholastic Aptitude Test) und der ACT (= American College Testing Program) dienen. Roznowski/Reith zeigen hierbei, dass einzelne mit Bias behaftete Items nicht unbedingt zu schlechten Gütemaßen der klassischen Testtheorie (= KTT) führen. Sie demonstrieren dies für Cronbachs Alpha als Reliabilitätsmaß, Korrelationen mit verschiedenen Außenkriterien als Maß für die Kriteriumsvalidität sowie für Korrelationen zwischen Teilskalen mit unterschiedlich starkem Bias als Maß für die innere Validität der Skalen.

Roznowski/Reith differenzieren daher zwischen Item Bias und Test Bias. Item Bias liegt dann vor, wenn nur einzelne Items verzerrt sind, im Hinblick auf den ganzen Test aber nur ein geringes Ausmaß an Verzerrung vorliegt. Test Bias hingegen bedeutet, dass der Test als Ganzes unfair ist und bestimmte Subpopulationen benachteiligt. Item Bias kann nach Auffassung von Roznowski/Reith ignoriert werden, Test Bias jedoch nicht. Wenn man dies auf die Situation in der Veränderungsmessung mit LLTMs und LLRAs verallgemeinert, so kann man den Begriff Test Bias mit starker Fehlspezifikation, den Begriff Item Bias mit schwacher Fehlspezifikation umschreiben. Aufgrund der Ergebnisse von [Baker, 1993] sollte man schwache Fehlspezifikation bei LLTMs jedoch nicht leichthin ignorieren, wie dies bei [Roznowski and Reith, 1999] geschieht. Wie in Kapitel 2.3 beschrieben, kann auch schwache Fehlspezifikation zu fehlerbehafteten Parameterschätzungen im LLTM führen.

3.2 Bias in der Veränderungsmessung

Zunächst definieren wir den Begriff „Bias in der Veränderungsmessung“. Wir nehmen dazu an, dass die Items eines (Teil-)Fragebogens alle denselben Trait messen. Unter dieser Voraussetzung sprechen wir von Bias bezüglich der Veränderung, wenn für ein Item in einer bestimmten Subpopulation ein anderer Veränderungsparameter angenommen werden muss als für den Rest der Items in dieser Subpopulation. Diese Definition von Bias in der Veränderung beschreibt somit, dass die Modellspezifikation zumindest für ein Item in einer Subpopulation nicht gültig ist.

3.2.1 Entdeckung von Fehlspezifikation mit Hilfe der logistischen Regression

In diesem Abschnitt entwickeln wir eine Methode zum Erkennen von Bias in der Veränderungsmessung. Ausgangspunkt dabei ist die Mantel-Haenszel-Technik zur Entdeckung von Item-Bias, wie sie z.B. in [Meredith and Millsap, 1992] für das Rasch-Modell dargestellt wird. Diese nutzt die Tatsache aus, dass gewisse Odds-Ratios konstant über alle Subpopulationen bleiben müssen, wenn der gemessene Schwierigkeitsparameter eines Rasch-Modells in allen Subpopulationen gelten soll.

Im Folgenden benutzen wir [Meredith and Millsap, 1992] als Ausgangspunkt und zeigen, dass das Odds-Ratio für die Lösungswahrscheinlichkeit eines Items i bezüglich zwei Messzeitpunkten unter gewissen Voraussetzungen konstant ist. Durch die Überprüfung dieser Eigenschaft des LLTMs kann festgestellt werden, ob zu einem bestimmten Item der angenommene Veränderungswert zwischen zwei Zeitpunkten zutreffen kann.

Zur Überprüfung dieser Eigenschaft des LLTMs kann ein (verallgemeinerter) Mantel-Haenszel-Test verwendet werden. Da jedoch der verallgemeinerte Mantel-Haenszel-Test über die Konstanz verschiedener Odds-Ratios in einer 3-dimensionalen Kontingenztafel ist (vgl. z.B. die Darstellung in [SAS Institute Inc., 1999]), kann man statt des Mantel-Haenszel-Tests auch äquivalente Testprozeduren aus der logistischen Regression verwenden. Letztere Modelle werden von uns bevorzugt, da sie besser verallgemeinert werden können. Im Allgemeinen bleibt jedoch zu sagen, dass die von uns vorgeschlagenen Methoden prinzipiell sowohl mit χ^2- und Likelihood-Ratio-Tests der logistischen Regression, als auch mit Hilfe der Mantel-Haenszel-Technik überprüft werden können.

Die hier dargestellte Methodik baut auf Theorem 13 aus [Meredith and Millsap, 1992] auf, in dem die Zulässigkeit des Mantel-Haenszel-Ansatzes für logistische Latent-Trait-Modelle und einen Messzeitpunkt bewiesen wird. Wir erweitern dieses Ergebnis auf die Veränderungsmessung mittels LLTM oder LLRA und erhalten somit ein Maß für die Zulässigkeit der modellierten Veränderung bei einem Item in einer bestimmten Subpopulation.

Für diese Methode setzen wir voraus, dass m Items an allen Zeitpunkten dargeboten werden, so dass $m_t = (t-1)m$ die Zahl der Items ist, die vor dem Zeitpunkt dargeboten wurden. Zur Vereinfachung setzen wir weiterhin voraus, dass die Veränderung durch einen einzigen Parameter δ_{tg} darstellbar ist, wobei der Index g für die untersuchte Subpopulation steht. Im Zeitpunkt t werden daher die Items $m_t + 1, \cdots m_t + m$ mit Schwierigkeiten $\beta_1 - \delta_{tg}, ..., \beta_m - \delta_{tg}$ beantwortet.

Zunächst berechnen wir die Wahrscheinlichkeit, dass ein beliebiges Item eines LLTMs in einer beliebigen Untersuchungsgruppe und zu einem beliebigen Zeitpunkt gelöst wird, falls der Summenscore R_v^t einer Person v im Zeitpunkt t den Wert r annimmt.

$$\begin{aligned} &Pr(X_{vlt} = 1 \mid R_v^t = r) = q_{t,l}(r) \\ &= \frac{\exp(\delta_{tg} - \beta_i)\gamma_{r-1}^l(\xi_{m_t+1,g}, ..., \xi_{m_t+m,g})}{\exp(\delta_{tg} - \beta_i)\gamma_{r-1}^l(\xi_{m_t+1,g}, ..., \xi_{m_t+m,g}) + \gamma_r^l(\xi_{m_t+1,g}, ..., \xi_{m_t+m,g})} \quad , \end{aligned} \tag{3.4}$$

wobei R_v der Summenscore einer Person v ist und

$$\xi_{(t-1)*m+i,g} = \exp(-\beta_i + \delta_{tg}) \tag{3.5}$$

gesetzt wird. $\gamma_r^l(\xi_{m_t+1,g}, ..., \xi_{m_t+m,g})$ ist die elementare symmetrische Funktion der Terme

$$\xi_{m_t+1,g}, ..., \xi_{m_t+l-1,g}, \xi_{m_t+l+1,g} ..., \xi_{m_t+m,g}$$

für einen Summenscore r im Zeitpunkt t.

Die Wettchance (= Odds) in einem beliebigen LLTM an einem beliebigen Zeitpunkt t erhält man mittels

$$\begin{aligned}&\frac{Pr(X_{vlt}=1 \mid R_v^t=r)}{Pr(X_{vlt}=0 \mid R_v^t=r)} = \\ &= \frac{\exp(\delta_{tg}-\beta_i)\gamma_{r-1}(\xi_{m_t+1,g},...,\xi_{m_t+l-1,g},\xi_{m_t+l+1,g},...,\xi_{m_t+m,g})}{\gamma_r(\xi_{m_t+1,g},...,\xi_{m_t+l-1,g},\xi_{m_t+l+1,g},...,\xi_{m_t+m,g})} \quad . \end{aligned} \tag{3.6}$$

Dieses Verhältnis kann für jeden beliebigen Zeitpunkt t aufgestellt werden. Jetzt betrachten wir das mit Hilfe von (3.6) gebildete Odds-Ratio zwischen der Wahrscheinlichkeit, ein Item zum Zeitpunkt t_1 zu lösen und der Wahrscheinlichkeit, das gleiche Item zu einem späteren Zeitpunkt t_2 zu lösen. Die Wettchance an einem beliebigen Zeitpunkt t in Untersuchungsgruppe g berechnet sich gemäß (3.6) bei Gültigkeit des vorausgesetzten Modells als

$$\begin{aligned}&Odds(t,l,r) = \\ &= \frac{Pr(X_{vlt}=1 \mid R_v=r)}{Pr(X_{vlt}=0 \mid R_v=r)} \\ &= \frac{q_{t,l}(r)}{1-q_{t,l}(r)} \\ &= \frac{\exp(\delta_{tg}-\beta_l)\gamma_{r-1}^{(t-1)m+l}(\xi_{(t-1)m+1,g},...,\xi_{tm,g})}{\gamma_r^{(t-1)m+l}(\xi_{(t-1)m+1,g},...,\xi_{tm,g})} \\ &= \frac{\exp(r\delta_{tg})\left[\exp(\beta_l)\gamma_{r-1}^{(t-1)m+l}(\exp(-\beta_1),...,\exp(-\beta_m))\right]}{exp(r\delta_{tg})\left[\gamma_r^{(t-1)m+l}(\exp(-\beta_1),...,\exp(-\beta_m))\right]} \\ &= \frac{q_{1,l}(r)}{1-q_{1,l}(r)} = Odds(1,l,r) \quad . \end{aligned} \tag{3.7}$$

Als nächstes betrachten wir zwei Zeitpunkte t_1 und t_2. Für die Wettchancen dieser Zeitpunkte gilt daher

$$\frac{Odds(t_1,l,r_{t_1})}{Odds(t_2,l,r_{t_2})} = \frac{Odds(1,l,r_{t_1})}{Odds(1,l,r_{t_2})} = const. \quad . \tag{3.8}$$

Insbesondere gilt:

$$\frac{Odds(t_1,l,r)}{Odds(t_2,l,r)} = 1 \quad , \tag{3.9}$$

falls die geschätzten Veränderungsparameter für Item l und den untersuchten Zeitpunkt zutreffend sind.

Andernfalls gilt mit $\delta^*_{t_1 g}$ als dem wahren Veränderungsparameter in Zeitpunkt t_1 und Item l in Gruppe g, sowie $\delta^*_{t_2 g}$ als wahrem Veränderungsparameter für Zeitpunkt t_2, Item l und Gruppe g:

$$\frac{\frac{q_{t_1 l}(r_{t_1})}{1-q_{t_1 l}(r_{t_1})}}{\frac{q_{t_2 l}(r_{t_2})}{1-q_{t_2 l}(r_{t_2})}} = \frac{\exp(\delta^*_{t_1 g} - \delta_{t_1 g}) \times Odds(t_1, l, r_{t_1})}{\exp(\delta^*_{t_2 g} - \delta_{t_2 g}) \times Odds(t_2, l, r_{t_2})} \quad . \tag{3.10}$$

Falls

$$\delta^*_{t_1 g} - \delta_{t_1 g} \neq \delta^*_{t_2 g} - \delta_{t_2 g} \tag{3.11}$$

gilt, kann man für ein vorgegebenes Item l die Gültigkeit dieser Bedingung über ein logistisches Regressionsmodell überprüfen. Andernfalls könnte 3.10 auch unter der Alternativhypothese gelten. Voraussetzung (3.11) ist stets erfüllt, wenn man statt zweier beliebiger Zeitpunkte t_1, t_2 die Zeitpunkte $1, t_2$ wählt, und falls $\delta_{t_2 g} \neq 0$ angenommen werden kann. Wenn man Zeitpunkt 1 als Referenzzeitpunkt verwendet, setzt man $\delta_{1g} = 0$ für alle Untersuchungsgruppen g voraus. (3.11) ist genau dann erfüllt, wenn zum Zeitpunkt t_2 eine von 0 verschiedene Veränderung stattgefunden hat und die Nullhypothese „Kein Item Bias für Item l“ gilt.

Diese Ergebnisse können für die Untersuchung auf verzerrte Items ausgenutzt werden. Zunächst betrachten wir dazu ein Item l bei den Zeitpunkten $t_1 = 1$ und t_2. Bei m Items pro Zeitpunkt kann der Summenscore $m-1$ verschiedene Werte annehmen, wenn man die uninformativen Scorewerte m und 0 außer Acht lässt. Wenn der geschätzte Veränderungsparameter zutrifft, kann eine $m-1 \times 2 \times 2$-Kontingenztafel gebildet werden, mit den Randmerkmalen „Summenscore“, „Messzeitpunkt“ und „Antwort bei Item l“. Mittels eines Mantel-Haenszel-Tests (vgl. z.B. [SAS Institute Inc., 1999]) kann überprüft werden, ob in jeder 2×2-Ebene dieser Kontingenztafel das Odds-Ratio den Wert 1 annimmt. Das Odds-Ratio $\widehat{OR}_r$ der 2×2-Ebene für Summenscore r ist dabei eine Schätzung für das Odds-Ratio aus (3.9) bei Summenscore r:

$$\widehat{OR}_r = \frac{\frac{x^r_{lt_1 \cdot}}{m - x^r_{lt_1 \cdot}}}{\frac{x^r_{lt_2 \cdot}}{m - x^r_{lt_2 \cdot}}} \quad . \tag{3.12}$$

$x^r_{lt_1 \cdot}$ ist hier die Zahl der Personen mit Summenscore r an Zeitpunkt t_1, die Item l richtig beantwortet haben. Ein solcher Test überprüft, ob bei allen Scoregruppen pro Zeitpunkt

der vorgegebene Veränderungsparameter vorliegen kann. Statt des Mantel-Haenszel-Tests kann auch ein Logit-Regressionsmodell verwendet werden, um die Konstanz der Odds-Ratios zu überprüfen.

Die beschriebene Vorgehensweise kann verallgemeinert werden. Wenn man mehrere Items gleichzeitig betrachten will, verwendet man die in [Fischer, 1995c] beschriebenen Techniken. Hier wird beschrieben, wie man die Mantel-Haenszel-Methode für das Rasch-Modell auf die Analyse von zwei Items gleichzeitig verallgemeinern kann. Die in Fischers Artikel verwendeten Argumente können ohne Weiteres auf das LLTM in der Veränderungsmessung übertragen werden.

Eine Untersuchung mehrerer Zeitpunkte gleichzeitig ist leicht möglich, wenn man die Daten für $T > 2$ Zeitpunkte in einer $m - 1 \times T \times 2$-Kontingenztafel organisiert. Bedingung (3.9) gilt hier genau dann, wenn diese Kontingenztafel unabhängig von der Variable „Messzeitpunkt“ ist, d.h. wenn ein loglineares Modell

$$EX_{rtj} = u + u_r + u_j + u_{rj} \tag{3.13}$$

gilt. Hierbei ist u ein Effekt für den Gesamtmittelwert, u_r der Effekt der Randvariable „Summenscore“, u_i der Effekt der Randvariable „Item l gelöst“ sowie u_{rj} die Interaktion zwischen den beiden Randvariablen. Hierbei haben die üblichen Normalisierungsbedingungen für loglineare Modelle zu gelten, also

$$\sum_r u_r = 0 \tag{3.14}$$

$$\sum_j u_j = 0 \tag{3.15}$$

$$\sum_r \sum_j u_{rj} = 0 \ . \tag{3.16}$$

Alternativ zum loglinearen Modell kann die Unabhängigkeit vom Messzeitpunkt auch mittels eines verallgemeinerten Mantel-Haenszel-Tests überprüft werden (vgl. zu diesem Test auch [SAS Institute Inc., 1999]).

Die vorgeschlagene Untersuchungsprozedur ist folgendermaßen motiviert: (3.8) gilt, wenn für Item l der gleiche Veränderungsparameter wie für die anderen Items angenommen werden kann. In diesem Fall ist die Wettchance $odds(q_{tl}(r))$ des LLTMs unabhängig vom Zeitpunkt t, und es ergibt sich bei jedem Zeitpunkt t und vorgegebenem Summenscore r_t die gleiche Wettchance. Diese Konstanzeigenschaft gilt auch für das Log Odds der beobachteten Lösungswahrscheinlichkeiten. Falls (3.8) zutrifft, dürfen die Log Odds nicht vom Messzeitpunkt T als zusätzlicher Variable abhängen. Falls der Messzeitpunkt einen signifikanten Einfluss auf die Log Odds besitzt, muss für das Item l ein anderer

Veränderungsparameter angenommen werden als für die übrigen Items. Die Zulässigkeit dieser Erweiterung im Zeitbereich folgt aus der im LLTM vorausgesetzten lokalen stochastischen Unabhängigkeit: Bei gegebenen Summenscores für zwei Zeitpunkte t_1 und t_2 kann Unabhängigkeit zwischen $X^l_{gt_1}$ und $X^l_{gt_2}$ angenommen werden.

Schließlich ist auch eine Erweiterung auf $G > 1$ Subpopulationen möglich: Das für $T > 2$ beschriebene loglineare Modell kann um eine zusätzliche Variable „Untersuchungsgruppe“ erweitert werden, so dass man für eine $(m-1) \times T \times G \times 2$-Kontingenztafel das Modell

$$EX_{rtgj} = u + u_r + u_j + u_{rj} \tag{3.17}$$

überprüft. Falls keine Fehlspezifikation vorliegt, kann die $(m-1)\times T\times G\times 2$-Kontingenztafel durch dieses Modell beschrieben werden.

3.2.2 Anwendungsbeispiel: Erlernen von syllogistischen Schlussweisen

Im Folgenden wird die zweite Modellierung anhand eines Datensatzes zum Erlernen syllogistischen Schließens ([Meiser, 1996]) demonstriert (vgl. auch Tabelle 3.1). Der Datensatz beschreibt die Fähigkeit von Kindern zum syllogistischen Schließen in Abhängigkeit von der Art des syllogistischen Schlusses und des Lebensalters, an dem ein solcher Schluss beobachtet wurde. Unterschieden werden die Syllogismusformen „Experiential Syllogism“ (d.h. durch Erfahrung erlernter Syllogismus) und „Abstract Syllogism“ (d.h. ohne Erfahrung aufgetretener syllogistischer Schluss). Beobachtet werden diese Items zu zwei verschiedenen Zeitpunkten, nämlich im Alter von 12 Jahren und im Alter von 15 Jahren.

Um diese Daten auf Bias bezüglich der Veränderung zu untersuchen, werden die Daten mit Hilfe der SAS-Prozedur FREQ analysiert. Es wird getrennt für die Items „Experiential Syllogism“ und „Abstract Syllogism“ ein Mantel-Haenszel-Test durchgeführt.

Ein einfaches Modell für diese Daten könnte lediglich einen Veränderungsparameter für beide Items annehmen. Dies würde bedeuten, dass sich die beiden Items auf sehr ähnliche Weise verändern. Falls die Nullhypothese des Mantel-Haenszel-Tests bei einem Item abgelehnt wird, muss angenommen werden, dass diese einfache Parametrisierung nicht geeignet ist, um die Daten zu beschreiben. Die Ergebnisse der Tests sind in der Tabelle 3.2 dargestellt.

Wie man sieht, können die Antwortmuster des Items „Abstract Syllogism“ vermutlich durch einen einzigen Veränderungsparameter erklärt werden. Für das Item „Experiential Syllogism“ ist dies nicht möglich. Dies deckt sich mit einem früheren Befund zu diesem Datensatz. [Meiser, 1996] untersucht diesen Datensatz mit einer loglinearen Repräsentation des Rasch-Modells. Die Parameterschätzung erfolgt hier mittels PROC CATMOD, einer SAS-Prozedur zur Analyse von Kontingenztafeln. Das von ihm angewendete

SyE, 12 Jahre	SyA, 12 Jahre	SyE, 15 Jahre	SyA, 15 Jahre	Häufigkeit
0	0	0	0	42
0	0	0	1	2
0	0	1	0	8
0	0	1	1	9
0	1	0	0	4
0	1	0	1	2
0	1	1	0	1
0	1	1	1	8
1	0	0	0	6
1	0	1	0	3
1	0	1	1	3
1	1	0	0	2
1	1	0	1	1
1	1	1	0	1
1	1	1	1	9

Tabelle 3.1: Datensatz aus [Meiser, 1996]. Abkürzungen: *SyE* = Experiential Syllogism, *SyA* = Abstract Syllogism.

Item	Freiheitsgrade	Testgröße	P-Wert
Abstract Syllogism	1	3.1983	0.0737
Experiential Syllogism	1	16.2784	< 0.0001

Tabelle 3.2: Ergebnisse des Mantel-Haenszel-Tests auf Fehlspezifikation

Modell schätzt für jedes Item eine separate Veränderung. Dabei ergibt sich eine positive Veränderung für Item SyE (Schwierigkeit -0.077 im ersten Messzeitpunkt vs. Schwierigkeit 0.408 im zweiten Messzeitpunkt) und eine negative Veränderung für Item SyA (Schwierigkeit 0.146 im ersten Messzeitpunkt vs. Schwierigkeit -0.547 im zweiten Messzeitpunkt).

Dieses unterschiedliche Verhalten bezüglich der Veränderung wird auch durch Signifikanztests belegt: Die zugehörige Likelihood-Ratio-Test-Statistik ergibt für das vollständige, wie oben aufgeführte Modell den Wert 7.08 bei 5 Freiheitsgraden und führt nicht zur Ablehnung des Modells bei Signifikanzniveau 0.05. Für ein einfacheres Modell mit konstanter Veränderung pro Person ergibt sich als Testgröße der Wert 17.40 bei 8 Freiheitsgraden, was zur Ablehnung des Modells bei Signifikanzniveau 0.05 führt. Bei einem direkten Vergleich zwischen vollständigem, wie oben aufgeführten Modell und dem vereinfachten Modell ergibt sich die Testgröße 10.32 bei 3 Freiheitsgraden, was ebenfalls zur Ablehnung des vereinfachten Modells bei Signifikanzniveau 0.05 führt.

Wie man sieht, kommt die von uns verwendete Methode im Wesentlichen zu den gleichen Ergebnissen wie die von [Meiser, 1996]: Die Annahme einer itemunabhängigen Veränderung muss aufgegeben werden. Unsere Methode verwendet ein einfacheres Modell, liefert aber die Zusatzinformation, welche Items mit einem eigenen Veränderungsparameter modelliert werden sollten. Dies kann vor allem von Nutzen sein, wenn mehr als 2 Items untersucht werden. Ein vorgeschalteter Test zur Untersuchung auf Bias bezüglich der Veränderung kann dann zu einem sparsameren Modell führen. Die Modellierung der Daten mit einem Rasch-Modell kann dadurch aber nicht ersetzt werden: Unsere Methode liefert weder Schwierigkeits-, noch Fähigkeits-, noch Veränderungsschätzungen.

Von Vorteil ist diese Methode vor allem bei der Untersuchung auf Fehlspezifikation, da sie ohne Kenntnis der geschätzten Parameterwerte verwendet werden kann. Geschätzte Parameterwerte können, wie die Untersuchungen von [Klein, 1999] und [Baker, 1993] zeigten, stark von den wahren Parameterwerten abweichen, und somit zu falschen Schlüssen führen.

Kapitel 4

Messung des Personenfits

In diesem Kapitel werden Methoden zur Messung des Personenfits vorgestellt. Im ersten Unterabschnitt werden allgemeine Voraussetzungen diskutiert, die ein Verfahren zur Messung von Personenfit erfüllen sollte.

4.1 Einsatz von Personenfitmaßen bei Item-Response-Modellen

Bei der Testkonstruktion mit Item-Response-Modellen (= IRT-Modellen) tauchen manchmal Antwortmuster auf, die mit den Voraussetzungen des Modells nur schwer in Einklang zu bringen sind. Beispiel: Eine Versuchsperson löst eine schwierige Aufgabe, obwohl ihre Fähigkeit sehr niedrig ist. Falls solche Fälle häufig auftreten, ist dies ein Anzeichen für eine mangelnde Übereinstimmung zwischen gewähltem IRT-Modell und angefallenen Daten. Es kann dann sinnvoll werden, Personen mit zu schlechtem Fit aus der Stichprobe zu entfernen, bzw. herauszufinden, aus welchen Gründen der Personenfit schlecht ist. Häufig angewandte Vorgehensweisen bei schlechtem Personenfit sind:

- Entfernen schlecht angepasster Personen aus der Kalibrierungsstichprobe, vgl. dazu z.B. [Schmitt et al., 1993].
- Einbeziehung einer Klasse unskalierbarer Personen, vgl. [Rost and v. Davier, 1995], [Rost, 1996].

Aus statistischer Sicht mutet das Entfernen schlecht angepasster Personen seltsam an, da die Personen Teil der Stichprobe sind, an der ein psychologischer Test kalibriert werden soll. Ein Entfernen solcher Probanden aus der Stichprobe kann den Anschein einer Manipulation der Stichprobe erwecken.

[Klauer, 1995] argumentiert gegen diesen Einwand, indem er die Struktur der Zufallsmechanismen in Betracht zieht, auf denen ein Item-Response-Modell aufbaut. Basierend auf [Holland, 1990] wird dabei dem Random Sampling Rationale, welches die Zufälligkeit auf das Ziehen einer Stichprobe zurückführt, das sogenannte Stochastic Subject Rationale entgegengesetzt. Dieses postuliert einen zufälligen Mechanismus innerhalb jeder Person, der dann in Abhängigkeit von Fähigkeit und Itemschwierigkeit eine richtige oder falsche Antwort einer Person bewirkt. Bei Annahme eines solchen Mechanismus ist es gerechtfertigt, ungeeignet antwortende Probanden auszusortieren, da sie sich (willentlich) nicht an das vereinbarte Testprocedere gehalten haben und somit die Stichprobe nur verfälschen würden. Die Vorteile der Messung des Personenfits lassen sich folgendermaßen begründen:

- Durch das Aussortieren von Personen, die sich nicht vereinbarungsgemäß mit dem Test beschäftigt haben, erhält man eine homogenere Stichprobe, und somit genauere Schätzungen der Itemparameter.
- Weiterhin kann man durch die Analyse der aussortierten Antwortmuster Regelmäßigkeiten entdecken, die zur Eliminierung schlecht passender Itemkombinationen führen können.

Problematisch an der Philosophie der Untersuchung des Personenfits bleibt jedoch, dass schlecht passende Antwortmuster zufällig auftreten können, wenn eine Person gemäß des zugrunde liegenden Modells antwortet (vgl. auch [Molenaar and Hoijtink, 1990]).

Daher ist es wichtig, einen Mechanismus zu finden, mit dem das zufällige Auftreten solcher Muster ausgeschlossen werden kann. Im Aufbau der meisten Personenfitindizes wird der Index bezüglich der Nullhypothese „Aufgetretenes Muster einer Person stammt aus dem Rasch-Modell“ auf Signifikanz geprüft. Mit Hilfe einer solchen Testprozedur kann man allerdings nur erkennen, ob ein Antwortmuster zu den $\alpha \times 100$ % seltensten Antwortmustern gehört; es kann keine Aussage darüber gemacht werden, ob dieses Antwortmuster durch das angenommene Item-Response-Modell oder durch einen anderen Mechanismus erzeugt wurde.

Um eine solche Aussage zu ermöglichen, kann man die Häufigkeit des Auftretens aberranter Muster in der Stichprobe mit der (unter dem postulierten IRT-Modell) anzunehmenden Häufigkeit solcher Muster vergleichen. Nur falls die Häufigkeit aberranter Muster in der Stichprobe den erwarteten Wert wesentlich übersteigt, kann man von einem Nachweis aberranter Antwortmechanismen ausgehen. Dies kann durch einen Test auf überzufälliges Vorkommen aller Aberranzen zusammen gewährleistet werden. Benötigt wird eine Methode zur Verhinderung der Akkumulation des Fehlers erster Art. Falls dies nicht berücksichtigt wird, und in einer Stichprobe k Tests zum Niveau α durchgeführt werden, ist der gesamte Fehler erster Art größer als α. Dann entscheidet man insgesamt zu häufig auf das Vorkommen aberranter Muster.

[Molenaar and Hoijtink, 1990] betonen, dass die geringe Auftretenswahrscheinlichkeit eines Antwortmusters unter dem postulierten Modell alleine kein Grund ist, dieses Ant-

wortmuster als aberrant zu bezeichnen, da ein schlecht angepasstes Muster auch zufällig auftreten kann. Ein Test auf Anpassungsgüte eines Musters untersucht hingegen nur, wie wahrscheinlich ein Muster auftritt. Seltene Muster werden als aberrant klassifiziert. Aus diesem Grund sollte bei jeder Untersuchung auf Personenfit bestimmt werden, ob aberrante Muster wirklich überzufällig häufig in der Gesamtstichprobe auftauchen. Falls dies der Fall ist, kann man gesichert auf mangelhafte Anpassung der Stichprobe schließen. [Klauer, 1991b] gibt ähnliche Empfehlungen.

4.2 Klassische Ansätze zur Messung des Personenfits

Likelihoodbasierte Ansätze zur Untersuchung des Personenfits

Likelihoodbasierte Ansätze gründen auf der Idee, dass man aus der Wahrscheinlichkeit, mit der ein Muster unter dem postulierten IRT-Modell auftritt, ein sinnvolles Maß für das Erkennen aberranter Items entwickeln kann: Je kleiner die Wahrscheinlichkeit für ein bestimmtes Antwortmuster ist, desto schlechter ist dieses Muster an das zugrunde gelegte Modell angepasst.

Der meistzitierte Ansatz auf diesem Gebiet wurde von Levine und Drasgow entwickelt (vgl. [Levine and Drasgow, 1988], [Levine and Drasgow, 1982], [Drasgow and Levine, 1986]). In diesen Ansätzen wird die Loglikelihood $L(\theta, \beta)$ eines Antwortmusters in Abhängigkeit von den vorgegebenen Schätzungen von Item und Personenparameter berechnet. Dabei nimmt man an, dass der Itemparameter β gegeben ist bzw. ausreichend genau geschätzt wurde. Mit dieser Voraussetzung ist die Loglikelihood eines Antwortmusters eine einparametrige Funktion des Fähigkeitsparameters θ. Diese Statistik ist in der Literatur als l_0 bekannt (vgl. [Levine and Drasgow, 1982], [Meijer and Sijtsma, 2001]):

$$l_0 = \sum_{i=1}^{m} \left[X_i \ln P\left(X_i = 1|\theta\right) + \left(1 - X_i\right) \ln\left(1 - P\left(X_i = 1|\theta\right)\right)\right] \ . \tag{4.1}$$

l_0 ist ein Maß für die Anpassungsgüte eines Antwortmusters, das bei schlechter Anpassungsgüte niedrige Werte annimmt. Da der Bereich niedriger Werte in l_0 aber mit θ schwankt, muss l_0 standardisiert werden. Somit erhält man den standardisierten Index

$$l_z = \frac{l_0 - El_0}{\sqrt{Var\left(l_0\right)}} \tag{4.2}$$

(vgl. z.B. [Meijer and Sijtsma, 2001]).

l_0 benötigt ein parametrisches Item-Response-Modell als Basis, um $P\left(X_i = 1|\theta\right)$ bestimmen zu können. Welches Modell dazu letztendlich verwendet wird, bleibt dem Anwender überlassen. Die Literatur nennt häufig das 3-parametrige logistische Modell (z.B. [Levine and Drasgow, 1982], [Schmitt et al., 1993], [Nering, 1995], [Meijer, 1997]), bzw. das 2-parametrige logistische Modell (vgl. hierzu [van Krimpen-Stoop and Meijer, 1999],

[Reise, 2000], [Reise and Widaman, 1999]). Da dieser Index hinreichend genaue Schätzungen für die Werte der Modellparameter benötigt, ist sein Einsatzbereich auf Fälle beschränkt, in denen die Modellparameter durch eine Kalibrierungsstudie bekannt sind, oder durch eine große Stichprobe mit hoher Genauigkeit bestimmt werden können.

Verschiedene Quellen beschäftigen sich mit der Frage, ob durch das Verwenden von Schätzungen für den Fähigkeitsparameter die Verteilungsannahmen ungültig werden. Mittels einer Simulationsstudie gelangt [Nering, 1995] zu dem Ergebnis, dass bei Schätzung des Fähigkeitsparameters die Normalverteilungsannehme für l_z nicht aufrecht erhalten werden kann. Dies trifft besonders auf kurze Tests mit weniger als 25 Items zu.

Ein auf besseren Verteilungsapproximationen beruhendes Testverfahren für Personenfit stellen [Molenaar and Hoijtink, 1990] vor. Sie schlagen in ihrer Arbeit die Likelihood eines Antwortmusters als Maß für den Personenfit vor, leiten approximative Verteilungen für die Likelihood eines Antwortmusters ab, und entwickeln einen darauf aufgebauten Test für Personenfit. Um eine bessere Verteilungsapproximaton auch für den Fall der Benutzung der CML zu erreichen, verwenden sie die Größe

$$M(x_v) = \sum_i \beta_i x_{vi} \tag{4.3}$$

als Testgröße. $M(x)$ weist die gleiche Ordnung bezüglich der Antwortmuster wie die Likelihood $L(x)$ auf, ist aber einfacher zu berechnen. Zudem kann für $M(x)$ leicht eine approximative Verteilung bestimmt werden, deren Approximationsgüte besser ist als die standardisierte Testgröße l_z.

Weiterhin untersuchen [Molenaar and Hoijtink, 1990] den Sinn eines Fitindex, der auf dem geschätzten Fähigkeitsparameter einer Person aufbaut. Sie weisen darauf hin, dass die Verwendung eines derartigen Personenfitmaßes zu Verzerrungen führt, da ein bestimmtes Antwortmuster aufgrund unterschiedlicher Fähigkeiten zustande kommen kann. Ein Antwortmuster kann zwar unter der geschätzten Fähigkeit der Person sehr unwahrscheinlich sein, unter einem anderen Fähigkeitswert jedoch durchaus plausibel. Abhilfe bei Problemen dieser Art kann durch das Verwenden der Conditional Maximum Likelihood (= CML) geschaffen werden. Von einer Verwendung der Marginal Maximum Likelihood raten Molenaar/Hoijtink wegen ungelöster Probleme mit der Verteilung dieser Likelihood hingegen ab. Bei Verwendung der CML als Grundlage des Personenfitindex können [Molenaar and Hoijtink, 1990] eine χ^2-Approximation der Verteilung ihrer Testgröße M(x) ableiten.

[Bedrick, 1997] schlägt eine alternative Anpassungsmethode vor, die auf einer Edgeworth-Approximation beruht. Diese schneidet bei den von [Bedrick, 1997] gerechneten Beispielen i.d.R. genauso gut ab wie die Approximation aus [Molenaar and Hoijtink, 1990]. Weiterhin vergleicht Bedrick die Güte beider Ansätze bei geschätzten und bei bekannten Momenten. Hier schneiden für beide Verfahren die Methoden mit geschätzten Momenten genauso gut ab wie die Methoden mit bekannten Momenten.

[Snijders, 2001] (vgl. dazu auch [van Krimpen-Stoop and Meijer, 1999]) zeigt, dass die Varianz einer Personenfitstatistik kleiner wird, wenn geschätzte statt der wahren Fähigkeitswerte verwendet werden. [Snijders, 2001] gibt ebenfalls eine Normalverteilungsapproximation für eine Personenfitstatistik l_z^* an. Diese leitet sich aus l_z durch die Aufnahme verschiedener Korrekturterme ab. Zur Berechnung von l_z^* werden WML-Schätzer (= Weighted Maximum Likelihood-Schätzer) anstelle der ML-Schätzer für die Fähigkeit verwendet. [van Krimpen-Stoop and Meijer, 1999] vergleichen die Verteilung von l_z mit bekannten Fähigkeitsparametern sowie von l_z bei WML-geschätzten Fähigkeitsparametern mit der Verteilung von l_z^*. Bei bekannten Fähigkeitsparametern stimmen Erwartungswert und Varianz der simulierten Verteilungen mit der zu erwartenden Verteilung überein. Allerdings sind die simulierten Verteilungen linkssteil und leicht leptokurtisch. Bei der mit geschätztem Fähigkeitsparameter arbeitenden Variante von l_z unterscheiden sich Erwartungswert und Varianz bei kleinen Stichproben deutlich von den vorgegebenen Werten. Auch sind hier die Verteilungen linksschief und deutlich leptokurtisch. Bei großer Zahl von Items schneidet diese Variante allerdings besser ab. l_z^* liegt zwischen den beiden gerade vorgestellten Varianten, schneidet allerdings bei kleinen Stichproben deutlich besser ab als l_z mit geschätzten Parameterwerten.

χ^2-basierte Ansätze zur Messung des Personenfits

Neben den likelihoodbasierten Ansätzen existieren auch einige χ^2-basierte Ansätze zur Messung des Personenfits. Die wichtigsten dieser Ansätze stammen von Tatsuoka (vgl. [Tatsuoka, 1984], [Li and Olejnik, 1997]), und sollen hier kurz dargestellt werden.

Tatsuokas Personenfitkoeffizienten sind stets als Quotient zweier Kovarianzen $Cov(x_v, z_1)$ und $Cov(z_2, z_3)$ definiert, wobei z_1, z_2, z_3 je nach Index variieren. Insgesamt leitete Tatsuoka sechs verschiedene Koeffizienten ab. Als Beispiel seien die Indizes ECI_2 und ECI_4 angegeben. Diese sind als die wichtigsten seiner Koeffizienten anzusehen (vgl. dazu auch [Li and Olejnik, 1997]). ECI_2 setzt $z_1 = G$, $z_2 = P_v$ und $z_3 = z_1 = G$, wobei

- P_v der Vektor der Antwortwahrscheinlichkeiten der Person v ist, sowie
- G der Vektor der mittleren Antwortwahrscheinlichkeiten einer Untersuchungsgruppe ist.

Der Koeffizient ECI_2 ergibt sich dann als

$$ECI_2 = 1 - \frac{Cov(x_v, G)}{Cov(P_v, G)} \quad . \tag{4.4}$$

[Tatsuoka, 1984] gibt eine standardisierte Form dieses Index an, auf die wir hier jedoch nicht näher eingehen.

Der Index ECI_4 ist nach dem gleichen Konstruktionsprinzip aufgebaut, verwendet jedoch $z_1 = z_2 = P_v$ und $z_3 = G$. Als Index erhält man dann:

$$ECI_4 = 1 - \frac{Cov(x_v, P_v)}{Cov(P_v, G)} \quad . \tag{4.5}$$

Auch zu diesem Index existiert eine standardisierte Variante. Das Funktionsprinzip dieser Indizes ist leicht erklärt: Die Kovarianz im Zähler des Bruchs (z.B. zwischen x_v und P_v misst den Zusammenhang zwischen Datenvektor und einer Zielgröße. Die Kovarianz im Nenner ist ein Normierungsfaktor bzw. eine (willkürlich gewählte) Vergleichsgröße, die bewirkt, dass der Wert des Bruchs immer zwischen 0 und 1 liegt. Bei perfekter Anpassung an den Normierungsfaktor erhält man für diese Indizes einen Wert, der bei 0 liegt. Hohe Werte der Indizes bedeuten schlechte Anpassung. ECI_2 ist als gruppenorientierter Index definiert, d.h. man vergleicht den Antwortvektor x_v einer Person mit der mittleren Antwortwahrscheinlichkeit in einer Gruppe. EC_4 ist dagegen als Individualindex konzipiert (vgl. [Li and Olejnik, 1997]).

Einfluss von Personenfit auf die Validität

Verschiedene Untersuchungen beschäftigen sich mit den Auswirkungen von Personenfituntersuchungen auf die Validität eines Tests. [Schmitt et al., 1993] berichten von sehr niedrigen Auswirkungen auf die Validität eines Tests. Wegen einiger methodischer Mängel (u.a. Verwendung von Speed-Tests) kann die Studie jedoch nicht verallgemeinert werden. [Meijer, 1997] erweitert die Ergebnisse dieser Studie, ohne deren methodischen Fehler zu wiederholen. Meijer weist nach, dass mangelnder Personenfit Auswirkungen auf die Validität eines Tests hat. Der Einsatz von Personenfitmaßen zur Entfernung schlecht angepasster Personen bewirkt jedoch keine Erhöhung der Validität. Dies ist nach Auffassung von Meijer vor allem darauf zurückzuführen, dass nur jeweils 60% der schlecht angepassten Personen überhaupt erkannt wurden. Der geringe Einfluss von Personenfitmaßnahmen auf die Validität eines Tests ist demzufolge ein Problem der geringen Teststärke der Personenfitindizes.

[Meijer, 1998] untersucht an einer Stichprobe von 410 Personen, ob Personenfitindizes hilfreich bei der Vorhersage des Verhaltens durch einen kognitiven Test sind. Als Test wird dabei der Verbal Analogies Test (= VAT) verwendet. Das Kriterium, an dem die Vorhersagequalität des Tests gemessen wird, ist die Einschätzung des Verhaltens der Versuchspersonen durch Beobachter bei der Arbeit in kleinen Gruppen von je 4 Personen. [Meijer, 1998] stellt in dieser Studie deutliche Unterschiede hinsichtlich der Vorhersagbarkeit des Verhaltens in Abhängigkeit vom Personenfit fest. Das Verhalten von Personen mit niedrigem Fit kann wesentlich schlechter vorhergesagt werden als das Verhalten von Personen mit hohem Fit. Zudem kann in der Gruppe mit schlechtem Personenfit eine wesentlich niedrigere Validität als in der gesamten Stichprobe festgestellt werden. Das Entfernen dieser Personen führt allerdings nur zu einem geringen Anstieg der Validität.

[Ferrando and Chico, 2001] diskutieren die Auswirkungen des Einsatzes von Personenfituntersuchungen auf die Validität eines Persönlichkeitstests. Hier wird untersucht, ob durch den Einsatz von Personenfittests Personen, die ein bestimmtes Verhalten vortäuschen, genauso gut erkannt werden können wie mit klassischen Methoden (z.B. eine Lügenskala oder eine Skala sozialer Erwünschtheit). Hierbei wurden in der Persönlichkeitsforschung übliche Skalen verwendet, nämlich die Extraversions-, Neurotizismus- und Psychotizismusskalen des Eysenck Personality Questionaire Revisited. In dieser Untersuchung schneiden Personenfittests wesentlich schlechter ab als die klassischen Methoden. Als Gründe hierfür werden die geringe Teststärke der Personenfitindizes bei einer kleinen Itemanzahl, ungüstige (d.h. zu niedrige) Werte des Trennschärfenparameters bei den untersuchten Skalen sowie der zu geringe Anteil von betrugsresistenten Items in den untersuchten Skalen angeführt.

[Schmitt et al., 1999] untersuchen u.a. den Zusammenhang zwischen mangelndem Personenfit und äußeren Variablen, wie z.B. Geschlecht und Ethnizität. Die hier aufgetretenen Korrelationen sind zwar signifikant von 0 verschieden, dies ist jedoch eher auf die von Schmitt et al. verwendete hohe Stichprobengröße von insgesamt 378 Versuchspersonen zurückzuführen. Die von [Schmitt et al., 1999] berichteten Korrelationskoeffizienten liegen zwischen 0.14 und 0.21 bei Geschlechtszugehörigkeit sowie zwischen -0.07 und -0.09 bei Ethnizität. Dies deutet zumindest bei den von Schmitt et al. verwendeten Skalen nicht auf starke Zusammenhänge zwischen schlechtem Personenfit und äußeren Merkmalen hin. Ebenso waren nur geringfügige Korrelationen zwischen den Personenfitmaßen der einzelnen Skalen zu entdecken. Weiterhin betonen die Autoren, dass sich mit Hilfe von Personenfitmaßen durchaus Personengruppen mit niedriger Validität entdecken lassen.

Zusammengefasst lässt sich sagen, dass die Entfernung von schlecht angepassten Personen nur geringe Auswirkungen auf die Testvalidität besitzt. Es scheint allerdings möglich zu sein, mit Hilfe von Personenfittests Subpopulationen mit niedriger Validität zu erkennen. Zudem kann u.U. die Vorhersage des Verhaltens einzelner Personen aufgrund von Personenfitmaßen besser beurteilt werden: Bei schlecht angepassten Personen scheint eine solche Vorhersage deutlich unsicherer zu sein als bei gut angepassten Versuchspersonen.

Personenfitkurven

Einen weitereren Ansatz zur Messung des Personenfits stellt die sog. „Person Response Function“ (= PRF) dar. Diese konstruiert in Analogie zur wohlbekannten Item-Response-Funktion eine Funktion, welche die Antwortwahrscheinlichkeit einer Person in Abhängigkeit von der Itemschwierigkeit modelliert. Auf [Lumsden, 1978] geht die Konstruktion von PRFs zurück. [Trabin and Weiss, 1983] schlagen den Einsatz im Rahmen der Personenfitmessung vor.

Zur Schätzung einer PRF sortieren [Sijtsma and Meijer, 2001] die Items nach ihrer (geschätzten) Schwierigkeit und fassen jeweils mehrere Items zu Gruppen zusammen. Innerhalb dieser Itemgruppen kann dann die Wahrscheinlichkeit geschätzt werden, dass ein Item dieser Schwierigkeitsgruppe richtig beantwortet wird. Aus den Antwortwahrscheinlichkeiten unter dem Item Response-Modell kann eine erwartete PRF konstruiert werden.

Dieser Ansatz eignet sich besonders gut dazu, Interaktionen zwischen Itemschwierigkeit und mangelndem Personenfit zu entdecken. Es existieren mehrere Möglichkeiten, um die Gleichheit von erwarteter und beobachteter PRF zu untersuchen. [Trabin and Weiss, 1983] schlagen einen χ^2-Anpassungstest vor. [Klauer and Rettig, 1990] schlagen einen Likelihood-Ratio-Test, einen Wald-Test und einen „Efficient Score"-Test vor (zu den Unterschieden zwischen diesen Testkonzepten: vgl. z.B. [Rao, 1973]).

Abgesehen von diesen generellen Testverfahren kommen auch Tests für spezielle Hypothesen zum Einsatz. So schlagen [Sijtsma and Meijer, 2001] zum Test auf ein monotones Fallen der PRF den Trendtest von Cochran-Armitage vor.

[Reise, 2000] benutzt Multidimensional Logistic Modelling (= MLM), um die Parameter eines Item Response-Modells und eine PRF gleichzeitig zu schätzen. Bei der MLM-Methodik handelt es sich um eine Erweiterung des normalen Regressionsmodells. Im MLM wird ein normales Regressionsmodell derart erweitert, dass zusätzliche Regressionsgleichungen eingeführt werden, in denen die Parameter des ursprünglichen Modells als abhängige Variable auftauchen. Dies sieht dann folgendermaßen aus (vgl. [Reise, 2000]):

$$\ln\left(\frac{P\left(X_{vi}=1\right)}{P\left(X_{vi}=0\right)}\right) = b_{0v} + b_{1v}\beta_i \tag{4.6}$$

$$b_{0v} = \gamma_{00} + \gamma_{01}\theta_v + \epsilon_{0v} \tag{4.7}$$

$$b_{1v} = \gamma_{10} + \epsilon_{1v} \tag{4.8}$$

Vorzugeben (durch Schätzung in einem Item Response-Modell) sind die Parameterwerte β_i und θ_v. Mittels der ersten Gleichungszeile wird eine logistische Funktion definiert, die die eigentliche Schätzung der PRF darstellt. Die zweite und dritte Zeile der obigen Gleichungen modellieren die Abhängigkeit der Formparameter b_{0v}, b_{1v} der logistischen Funktion von der Fähigkeit θ_v der Person v sowie von Fehlertermen ϵ_{0v},ϵ_{1v}. [Reise, 2000] kommt so zur Schätzung von genau einer logistischen Funktion pro Person. Diese Schätzungen können dann auf einen fallenden Trend oder ähnliche Hypothesen hin untersucht werden.

Weitere Erkenntnisse zu den klassischen Ansätzen

Hier wird auf weitere Erkenntnisse zu diesen klassischen Personenfitindizes eingegangen. Die Auswahl der vorgestellten Ergebnisse erhebt keinen Anspruch auf Vollständigkeit.

[Li and Olejnik, 1997] berichten über eine Simulationsstudie, die mehrere dieser klassischen Indizes (darunter standardisierte Formen von EC_2, EC_4 und Drasgows Likelihood-Index l_z) miteinander vergleicht. Dabei wurden folgende Ergebnisse festgestellt:

1. Die empirischen Verteilungen dieser Stichproben weichen signifikant von der angenommenen Normalverteilung ab.
2. Alle Indizes zeigen i.d.R. vergleichbare Erfolgsraten, unabhängig von der Zahl der Items und der Art des „Missfits“.
3. Je mehr Items vorliegen, desto besser funktionieren diese Indizes.
4. Es kann kein Zusammenhang zwischen den Fähigkeitsschätzwerten und den Werten der Indizes nachgewiesen werden.
5. Die zu den Indizes gehörenden Signifikanztests sind i.d.R. konservativ.

[Hornke and Habon, 1986] stellen eine Untersuchung vor, in der ein LLTM zusammen mit verschiedenen Personenfitindizes zur Konstruktion einer Itembank verwendet wird. Auf diese Weise soll ein regelgeleitetes Verfahren zur Konstruktion von Itembanken konstruiert werden. Dazu entwerfen die Autoren zunächst einen Satz von Regeln zur Itemkonstruktion. In dem von Hornke/Habon vorgestellten Datensatz handelt es sich um sog. Matrix-Items, bei denen miteinander zusammenhängende Symbole in einer 3×3-Matrix vorgegeben sind, wobei stets eine Zelle der Matrix nicht besetzt ist. Die Probanden sollen dann die freie Zelle dieser Matrix so besetzen, dass das Konstruktionsprinzip der Matrix nicht verletzt wird. Die verwendeten Symbole können in vorgegebener Weise variieren.

Je nach Wahl der Symbole und der Regeln, mit denen die Zellen untereinander zusammenhängen, werden unterschiedliche kognitive Operationen benötigt. Hornke/Habon verwenden Symbole, die sich in Form und Füllmuster voneinander unterscheiden, und kommen damit auf 5 relevante kognitive Operationen:

- Identifikation der Form: Parameter β_1
- Entdecken der Regel, nach der die Formen angeordnet wurden: Parameter β_2
- Identifikation der relevanten Füllmuster: Parameter β_3
- Entdecken der Regel, nach der die Füllmuster angeordnet wurden: Parameter β_4
- Zusammenfügen der Regeln: Parameter β_5.

Diese 5 Operationen können durch ein spezielles LLTM modelliert werden. In der Itembank von Hornke/Habon werden mit Hilfe von 8 Kompositionsregeln und 3 Methoden, mit denen diese Kompositionsregeln zusammenwirken können, 616 Items konstruiert. Anhand einer Stichprobe von 7400 Personen werden für jedes Item i die Parameter $\beta_{i1}, \beta_{i2}, ..., \beta_{i5}$ geschätzt. Der Fit des vorausgesetzten LLTM wird mittels einer Kombination von χ^2-basierten Item- und Personenfitmaßen überprüft.

Mit dieser Methode werden 134 schlecht fittende und somit nicht den Homogenitätsanforderungen des LLTM genügende Items entfernt. Mit Hilfe der Personenfitmaße kann festgestellt werden, dass der schlechte Fit mancher Items nur auf einige wenige Personen zurückzuführen ist, die bei diesen Items ein stark abweichendes Verhalten aufweisen. Durch Entfernung von 360 Versuchspersonen (also ca. 5 % der Stichprobe) kann ein ausreichender Fit an das postulierte Modell erreicht werden. Weiterhin kann durch das Zusammenspiel von Personenfit- und Itemfit erreicht werden, dass Ursachen für mangelhaften Itemfit erkannt werden. Dies wird dadurch ermöglicht, dass durch das Zusammenspiel von Item- und Personenfit nur solche Abweichungen vom postulierten Modell berücksichtigt werden, die klar bei einer Menge mehrerer Probanden identifiziert werden können. So berichten Hornke/Habon von Items, bei denen eine Subpopulation eine unterschiedliche Lösungsstrategie als der Rest der Population anwendet. Hornke/Habon demonstrieren hiermit, dass der Einsatz von Personenfitmaßen sinnvoll zur Entdeckung falsch spezifizierter LLTMs verwendet werden kann.

4.3 Optimale Tests zur Untersuchung des Personenfits

Optimale Signifikanztests für Personenfit maximieren bei gegebenem Signifikanzniveau α die Gütefunktion eines Tests über den ganzen Parameterraum gleichmäßig und minimieren somit die Wahrscheinlickeit eines Fehlers zweiter Art (weiterführende Werke zum Konstruktionsprinzip gleichmäßig bester unverfälschter Tests: z.B. [Witting, 1978], [Witting, 1985]; [Lehmann, 1997]).

Grundlegend für die Existenz solcher optimaler Tests für den Personenfit ist dabei, dass die Verteilungsfamilie der Likelihood eines Rasch-Modells sich als Exponentialfamilie von Verteilungen darstellen lässt. Dies führt aufgrund der Neyman-Pearsonschen Signifikanztest-Theorie (vgl. auch [Lehmann, 1997]; bzw. [Klauer, 1991b], [Klauer, 1995]; [Ponocny, 2000]) zur Existenz und Bestimmbarkeit sogenannter gleichmäßig bester unverfälschter Tests. In den Jahren 1991 bis 1995 veröffentlichte Klauer (z.B. [Klauer, 1991b], [Klauer, 1995]) einige Artikel, in denen er unter den Gesichtspunkten der frequentistischen Testtheorie gleichmäßig beste Tests für den Personenfit vorstellte (Anmerkung: „Optimale“ Tests werden hier als Synonym für „gleichmäßig beste“ Tests verwendet).

Der von Klauer vorgestellte Ansatz sieht als Alternativhypothese die Situation vor, in der

ein Teil der Items einen anderen Trait misst, so dass es für alle Probanden zwei Traits gibt: Den zu messenden Trait θ_v und den ungewollten Trait, dessen Fähigkeitswerte mit $\theta_v + \lambda_v$ bezeichnet werden. [Ponocny, 2000] erweiterte diesen Ansatz auf allgemeinere Alternativhypothesen.

Ausgangspunkt dieser Ansätze ist die Darstellung der Likelihood eines Rasch-Modells als Mitglied der Exponentialfamilie von Verteilungen, d.h. in folgender Form:

$$P(X_v = x_v \mid \theta, \beta) = \mu(\theta) h(x_v) exp[T(x_v)\theta] \tag{4.9}$$

mit:

$$\mu(\theta) = \left(\prod_i (1 + exp(\theta - \beta_i) \right)^{-1} \sum_{x \mid \sum x_i = r} exp(- \sum x_j \beta_j) \quad . \tag{4.10}$$

Wenn wir mit $R(X)$ den gesamten Summenscore bezeichnen sowie mit $T(X_2)$ den Summenscore der Variablen, die den zweiten Trait messen, erhalten wir in der Schreibweise als Element der Exponentialfamilie folgende Likelihood:

$$P(X_v = x_v) = \mu_1(\theta)\mu_2(\theta + \lambda) h_1(x_1) h_2(x_2) exp\left(R\left(X\right)\theta + T(X_2)\lambda\right) \quad . \tag{4.11}$$

Dabei ist x_1 (bzw. x_2) das Antwortmuster der zum intendierten Trait gehörenden (bzw. zum ungewollten Trait gehörenden) Items. Für diese zweiparametrige Exponentialfamilie von Verteilungen kann man jetzt optimale Tests zur Nullhypothese

$H_0 : \quad \lambda_v = 0$

durchführen. Gegenhypothese ist dann

$H_A : \quad \lambda_v \neq 0.$

Nach der Neyman-Pearsonschen Signifikanztest-Theorie kann in diesem Fall ein gleichmäßig bester unverfälschter Test konstruiert werden, der maximale Teststärke bezüglich der oben angeführten Gegenhypothese besitzt. [Klauer, 1991b], [Klauer, 1995] und (in der Fortsetzung dieser Artikel) [Ponocny, 2000] stellen somit ein System optimaler (frequentistischer) Tests vor, mit deren Hilfe die Anpassungsgüte eines Antwortmusters gegen bestimmte spezifische Verletzungen des Rasch-Modells getestet werden kann. Durch die Annahme spezifischer Gegenmodelle können (unspezifische) Anpassungstests für Antwortmuster auf parametrische Testprobleme reduziert werden, für die (im Sinne des verallgemeinerten Neyman-Pearson-Lemmas, siehe auch [Witting, 1985], Satz 2.67) optimale Testverfahren existieren. Durch die Beschränkung auf ein konkretes Gegenmodell können Begründungen für das schlechte Verhalten eines Antwortmusters gegeben werden (z.B. wie in [Klauer, 1991b]: Einzelne Items messen einen zweiten Trait).

Für die Probleme der Akkumulierung des Fehlers erster Art gibt Klauer folgende Lösung an: Es wird ein vorgeschalteter Test auf Signifikanz aller möglicherweise aberranten Antwortmuster vorgeschlagen, und nur wenn die Globalhypothese abgelehnt wird, gelangen Tests für einzelne Antwortmuster zum Einsatz. Dieses Vorgehen bewirkt, dass nur dann aberrante Muster ausgewiesen werden, wenn diese auch wirklich überzufällig auftreten.

[Ponocny, 2000] erweitert das Verfahren von Klauer auf ein allgemeineres Modell, das ähnlich wie die Klasse der loglinearen Rasch-Modelle (vgl. [Kelderman, 1984]) aufgebaut ist. Außerdem werden Algorithmen zur schnelleren Berechnung der exakten Verteilung der Testgröße des gleichmäßig besten unverfälschten Tests vorgestellt.

4.4 Mixed Rasch-Modelle und Bayes-Statistik

In diesem Abschnitt wird ein Vergleich der bisher vorgestellten Personenfitindizes mit den sog. Mixed-Rasch-Modellen durchgeführt, die von manchen Autoren (vgl. [Rost, 1999], S. 150) für die Erkennung von schlechtem Personenfit empfohlen werden.

Bei der bayesianischen Testtheorie liegt folgendes, der bayesianischen Diskriminanzanalyse entlehntes Konzept vor (vgl. [Falk et al., 1995], S. 207ff, [Krzanowski and Marriott, 1995] oder [Berger, 1985], Kap. 4.4.3. für eine allgemeinere Darstellung der Bayesianischen Entscheidungstheorie): Gegeben seien ein Beobachtungsvektor $\mathbf{X_v}$, dessen Verteilung von einem Parameter ξ abhängt, sowie eine Menge $H_1, H_2, \ldots, H_C$ von Hypothesen über diesen Parameter ξ. Weiterhin seien die A-Priori-Wahrscheinlichkeiten $\pi_1, \ldots, \pi_C$ bekannt, mit der diese Hypothesen auftreten. Die Wahrscheinlichkeit, dass eine Realisation $\mathbf{x_v}$ von $\mathbf{X_v}$ beobachtet wird, ergibt sich dann durch:

$$Pr(\mathbf{X_v} = \mathbf{x_v}) = \sum_i \pi_i Pr(\mathbf{X_v} = \mathbf{x_v}|H_i) \quad . \tag{4.12}$$

Weiterhin sei $G_1, \ldots, G_k$ eine (unbekannte) Partition des Stichprobenraums S. Schließlich liege eine Kostenfunktion $K(j,i) : \{1, \ldots, C\} \times \{1, \ldots, C\} \mapsto [0, \infty]$ vor, die die Kosten einer Entscheidung für die Hypothese H_j angibt, wenn in Wirklichkeit die Hypothese H_i vorliegt. Im Fall der Entscheidung für die richtige Hypothese sollen keine Kosten anfallen.

Gesucht ist eine Entscheidungsregel der Art

> Falls $\mathbf{X_v}$ in die Partition G_k fällt ($k = 1, \ldots, C$), so wird die Hypothese H_k gewählt.

Diese Entscheidungsregel soll das sog. Bayesrisiko

$$R = \sum_j \sum_i \pi_i K(j,i) Pr\left(\mathbf{X_v} \in G_i \mid H_j\right) \tag{4.13}$$

minimieren.

In dieser Situation existiert eine optimale Entscheidungsfunktion (vgl. [Falk et al., 1995]). Falls eine einfache symmetrische Kostenfunktion mit

$$K\left(j,i\right) = K$$

für alle $j \neq i$ gilt, so fällt die Wahl auf die Entscheidungsfunktion: Entscheidung für H_k falls

$$Pr(\mathbf{X_v} = \mathbf{x_v}) = \max_i Pr(\mathbf{X_v} = \mathbf{x_v} \,|H_i\,) \quad . \tag{4.14}$$

Das (dichotome) Mixed-Rasch-Modell modelliert die Wahrscheinlichkeit $Pr(\mathbf{X_v} = \mathbf{x_v})$ des Auftretens eines Antwortvektors $\mathbf{X_v}$ durch die Gleichung

$$Pr(\mathbf{X_v} = \mathbf{x_v}) = \sum_i \pi_c Pr(\mathbf{X_v} = \mathbf{x_v}|c) \quad , \tag{4.15}$$

wobei $c \in \{1, \ldots, C\}$ der Index der latenten Subpopulation c ist. Dabei ist $Pr(\mathbf{X_v} = \mathbf{x_v}|c)$ die Wahrscheinlichkeit des Antwortvektors $\mathbf{X_v}$ unter Gültigkeit eines Rasch-Modells mit den in Subpopulation c zutreffenden Parameterwerten (vgl. z.B. [Meiser et al., 1998]). Dabei sind die in den einzelnen Subpopulationen gültigen Parameterwerte β_i^c, $1 \leq i \leq I$ und $1 \leq c \leq C$ und die Auftretenswahrscheinlichkeiten der einzelnen Subpopulationen π_c aus der Stichprobe zu schätzen.

Wie man jetzt leicht sieht, erhält man (4.12), indem man $\hat{\pi}_c$ als A-Priori-Wahrscheinlichkeiten verwendet, sowie die in den einzelnen Subpopulationen geschätzten Parametervektoren $\left(\hat{\beta}_1^c, \ldots, \hat{\beta}_I^c\right)^T$ als Hypothese H_c.

Aufgrund dieser Interpretationsmöglichkeit eignet sich das Mixed-Rasch-Modell dazu, Aussagen über den Personenfit zu machen. Die Hypothesen H_c können als die zur Auswahl stehenden Subpopulationen interpretiert werden. Ziel ist dann die Auswahl einer Subpopulation bei gegebenem Antwortmuster. Grundlage ist hierbei eine konstante Kostenfunktion. Man wählt diejenige Subpopulation, durch die die Wahrscheinlichkeit (4.14) maximiert wird (vgl. dazu auch [Rost and v. Davier, 1995]). Dies ist eine sehr einfache und elegante Form, um schlechten Personenfit entdecken zu können (vgl. dazu auch [Rost, 1999], S. 150).

Die Arbeit mit dem Mixed-Rasch-Modell führt stets zu einer Auswahl zwischen (durch vorhergehende Schätzung) bekannten Subpopulationen. Eine Einteilung in gut und schlecht

angepasste Versuchspersonen ist nur möglich, wenn schlecht angepasste Subpopulationen in das geschätzte Modell mit aufgenommen wurden. Dies ist jedoch nicht immer der Fall, da die Zahl der aufgenommenen Subpopulationen im Ermessen des auswertenden Statistikers liegt. So berichten z.B. [v. Davier and Rost, 1995] von einer Untersuchung, bei der eine 2-Klassenlösung gefunden wurde: eine Klasse der skalierbaren Personen, und eine Klasse der „Unskalierbaren“ (vgl. auch [Rost and Georg, 1991]). [Meiser et al., 1998] verwenden eine (restringierte) 2-Klassenlösung, bei der in beiden Klassen unterschiedliche Modelle verwendet werden. Eine Klasse schlecht angepasster Versuchspersonen kommt bei diesen Autoren daher nicht vor.

Bei der Unterscheidung zwischen verschiedenen Subpopulationen wird im Mixed-Rasch-Modell der ganze Antwortvektor einer Versuchsperson verwendet. Eine Fokussierung auf einige wenige Parameter, wie z.B. einen Veränderungsparameter, scheint kaum möglich zu sein. Aus diesem Grund kann nach Ansicht des Autors das Mixed-Rasch-Modell nicht zum Erkennen von Fehlspezifikation in der Veränderungsmessung verwendet werden. Zudem liefert das Bayes-Konzept der statistischen Tests stets eine Wahrscheinlichkeit für das Zutreffen einer bestimmten Hypothese H_c. Dies ermöglicht aber in vielen Fällen keine klare Entscheidung bzw. kein gut interpretierbares Maß für die Abweichung von einem vorgegebenen Modell.

Kapitel 5

Personenfitmessung im Veränderungsmodell

In diesem Kapitel werden Verfahren zur Entdeckung von mangelndem Personenfit bei der Veränderungsmessung mit dem LLTM bei 2 Zeitpunkten vorgestellt. Dabei erklären wir zunächst, welche Hypothesen besonders berücksichtigt werden sollen. Dann wird eine Zusammenfassung der für das Verständnis notwendigen Aspekte der Neyman-Pearsonschen Testtheorie (vgl. für eine Gesamtdarstellung z.B. [Witting, 1978]) gegeben. Von dieser allgemeinen Theorie ausgehend, entwickeln wir schließlich, auf dem LLTM aufbauend, statistisch optimale Personenfittests für die Veränderungsmessung.

5.1 Nullhypothesen für Personenfit bei Rasch-Modellen der Veränderungsmessung

Ursprünglich wurden Personenfittests entwickelt, um aberrante Antwortmuster bei der Beantwortung psychologischer Tests zu entdecken und auszusortieren. Die meisten dieser Verfahren sind unspezifisch, d.h. gegen keine spezielle Alternativhypothese gerichtet (vgl. z.B. [Levine and Drasgow, 1982], [Meijer, 1994], [Meijer, 1995], [Li and Olejnik, 1997], [Meijer and Sijtsma, 2001]). [Levine and Drasgow, 1988] stellt eine Ausnahme von diesem Befund dar: hier sind spezielle Indizes gegen Abschreiben, Raten und ähnliche Probleme zu finden. Der Hauptteil der Literatur zu Personenfitmodellen verwendet nicht das Rasch-Modell oder eine seiner Erweiterungen, sondern das 2-parametrige resp. 3-parametrige logistische Modell. Optimalität für das Rasch-Modell ist bei diesen Tests nicht nachgewiesen.

Bei der für die Familie der Rasch-Modelle spezifischen Literatur werden i.d.R. spezielle Gegenmodelle berücksichtigt. [Molenaar and Hoijtink, 1990] verwenden in ihrer Arbeit Tests bezüglich eines Diskriminationsparameters ξ sowie eine unspezifische Hypothese, die gegen keine bestimmte Art der Aberration gerichtet ist. Somit kann der von Molenaar/Hoijtink

eingeführte Test auch zur Diskrimination zwischen 1- und 2-parametrigen logistischem Item Response-Modellen verwendet werden.

[Klauer, 1991b], [Klauer, 1991a], [Klauer, 1995] und [Ponocny, 2000] stellen Tests vor, deren Alternativhypothesen aus der Klasse der loglinearen Rasch-Modelle (zu letzteren: vgl. [Kelderman, 1984]) stammen. Letztere lassen sich mit Suffizienzargumenten auf Verteilungen aus der Klasse der einparametrigen Exponentialfamilien zurückführen. Wie [Klauer, 1995] zeigt, lassen sich die meisten Typen von Aberrationen mit Hilfe dieser Modellklasse modellieren. Klauer erwähnt u.a. Tests bezüglich der Eindimensionalität des Traits und bezüglich Interaktionen zwischen verschiedenen Items. Die von [Klauer, 1995], [Molenaar and Hoijtink, 1990] benannten Testprobleme können im LLTM genauso auftreten wie im normalen Rasch-Modell. Unserer Auffassung nach sollten diese Testprobleme daher auch auf die gleiche Weise behandelt werden, nämlich durch die in den Artikeln [Klauer, 1995], [Klauer, 1991b], [Klauer, 1991a] oder [Molenaar and Hoijtink, 1990] vorgeschlagenen Tests. Dies betrifft unserer Auffassung nach alle Fälle von mangelndem Personenfit, bei dem die Veränderung nicht betroffen ist. In diesen Fällen kann man das LLTM wie ein normales Rasch-Modell behandeln. Für diese Fälle werden daher im folgenden Abschnitt keine Verfahren vorgeschlagen.

Unserer Meinung nach eignen sich besonders die auf der PRF aufbauenden Verfahren für den Einsatz bei LLTMs. In der Veränderungsmessung werden i.d.R. sehr viele Items dargeboten, so z.B. bei zwei Zeitpunkten mit je 30 Items insgesamt 60 Items. Diese i.d.R. hohe Zahl an (virtuellen) Items erleichtert es, Itemgruppen zu bilden und daraus eine PRF zu konstruieren. Da meistens nicht alle Traits gleichmäßig auf die Schwierigkeitsgruppen verteilt sind, wird sich die PRF oft dazu eignen, allgemeine Aussagen über die Anpassungsgüte einer Person bezüglich eines bestimmten Traits zu machen.

Bisher wurden in dieser Arbeit bekannte Ansätze zur Messung des Personenfits vorgestellt. Im Folgenden stellen wir Erweiterungen der bekannten Ansätze vor, die auf eine Untersuchung der Veränderungsparameter zugeschnitten sind. Bevor wir auf technische Einzelheiten eingehen, fassen wir die neuen Verfahren kurz zusammen.

Im Rest des vorliegenden Kapitels beschäftigen wir uns mit der Frage, ob ein Veränderungsparameter bei einem LLTM mit zwei Zeitpunkten einen vorgegebenen Wert δ_0 annimmt. Für Hypothesen dieser Art stellen wir gleichmäßig beste Signifikanztests vor. Diese Tests empfehlen wir auch dann, wenn nur ein Teil der Items daraufhin untersucht werden soll, ob der Veränderungsparameter in dieser Teilmenge von Items den Wert δ_0 annimmt. Das gleiche Problem wird in Kapitel 8.1 behandelt, allerdings mit der Erweiterung auf Situationen, bei denen an den beiden Testzeitpunkten unterschiedliche Items verwendet werden. Kapitel 6.3.2 beschäftigt sich mit Tests für 2 Zeitpunkte, aber mehreren Traits. Untersucht werden dabei Hypothesen über die gleichzeitige Veränderung mehrerer Traits, also z.B. der Art:

H_0: $\delta_1 = c_1$ und gleichzeitig $\delta_2 = c_2$.

In Kapitel 6.3.1 werden Hypothesen bezüglich der Gleichheit zwischen mehreren Veränderungstraits behandelt. Diese nehmen z.B. die Form

$H_0 : \delta_1 = \delta_2$

an. Hypothesen dieser Art können eingesetzt werden, um die Existenz eines zusätzlichen Veränderungstraits zu überprüfen: Dazu nehmen wir an, dass für alle m Items der Trait δ_1 zutrifft. Für eine Teilmenge von m_1 Items kann ein anderer Trait δ_2 zutreffen. Um dies zu überprüfen, genügt ein Vergleich zwischen den m_1 Items mit möglicherweise verändertem Trait und den $m - m_1$ restlichen Items bezüglich obiger Hypothese.

Wechselwirkungen zwischen zwei Traits können durch den Einsatz von Hpothesen des Typs $\delta = \delta_0$ überprüft werden. Dazu betrachten wir nur die Items, bei denen die beiden interessierenden Traits gemeinsam vorkommen. Falls λ die Wechselwirkung dieser Traits beschreibt, sowie $\hat{\delta_1}$ resp. $\hat{\delta_2}$ die (geschätzten) Veränderungen der untersuchten Traits, so kann durch die Hypothese

H_0: $\delta = \hat{\delta_1} + \hat{\delta_2}$

überprüft werden, ob eine Wechselwirkung vorliegt. Auch Tests bezüglich Verletzungen der lokalen stochastischen Unabhängigkeit können durchgeführt werden. In diesem Zusammenhang interessieren uns Verletzungen der lokalen stochastischen Unabhängigkeit zwischen (virtuellen) Items, die an unterschiedlichen Zeitpunkten abgefragt wurden. Ein Beispiel für die Behandlung solcher Hypothesen wird in Kapitel 8.2 untersucht.

5.2 Die allgemeine Form gleichmäßig bester unverfälschter Tests

In diesem Kapitel stellen wir dar, welche Form gleichmäßig beste unverfälschte Tests in einparametrigen Exponentialfamilien besitzen. Daher folgt an dieser Stelle ein kleiner Exkurs in die Neyman-Pearsonsche Testtheorie, in dem die für unseren Zusammenhang wichtigsten Ergebnisse dieser Theorie erläutert werden.

Untersucht wird eine Zufallsgröße X, deren Verteilung zu der Klasse der einparametrigen Exponentialfamilien gehört. Signifikanztests werden in diesem Zusammenhang mit Φ bzw. Φ^* bezeichnet.

Die Verteilung einer Zufallsgröße X gehört zur Klasse der einparametrigen Exponentialfamilien, wenn sich ihre Dichte $f(x, \vartheta)$ in der folgenden Form darstellen lässt:

$$f(x, \vartheta) = g(\vartheta)\, h(x) \exp[\vartheta T(x)] \ . \tag{5.1}$$

$T(X)$ ist die suffiziente Statistik für einen interessierenden Parameter ϑ. Untersucht werden Nullhypothesen über diesen interessierenden Parameter ϑ. $v(\vartheta)$ wird auch als natürlicher Parameter der Exponentialfamilie bezeichnet. (Zur Klasse der Exponentialfamilien: vgl. z.B. [Witting, 1978]).

Aus dem verallgemeinerten Fundamentallemma von Neyman/Pearson folgt die Existenz und Eindeutigkeit gleichmäßig bester Tests über den Parameter ϑ (vgl. [Witting, 1978], Kap. 2.8.).

In m-parametrigen Exponentialfamilien sind $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_m)$ sowie $\boldsymbol{T} = (T_1(x), \ldots, T_m(x))$ vektorielle Größen. In einer m-parametrigen Exponentialfamilie lässt sich die Dichte $f(x, \boldsymbol{\vartheta})$ in der folgenden Form darstellen:

$$f(x, \boldsymbol{\vartheta}) = g(\boldsymbol{\vartheta})\, h(x) \exp\left[\sum_i \vartheta_i T_i(x)\right] \ . \tag{5.2}$$

Dabei sind die Vektoren $\boldsymbol{\vartheta}$ und $\boldsymbol{T}$ linear unabhängig voneinander. Die Funktionen $T_i(x)$ sind suffiziente Statistiken für die Parameter ϑ_i. Im Allgemeinen existieren in Verteilungen aus mehrparametrigen Exponentialfamilien keine besten Tests. Es gibt aber eine Ausnahme, nämlich den Fall „1 interessierender Parameter, m-1 störende Nuisance-Parameter“. Um dies näher zu erläutern, gehen wir davon aus, dass wir einen Test über den Parameter ϑ_i durchführen wollen. Die Parameter ϑ_1,..., ϑ_{i-1},ϑ_{i+1},...,ϑ_m sind in diesem Fall Nuisance-Parameter. Suffiziente Statistiken für die Nuisance-Parameter sind $T_1(x)$, ...,$T_{i-1}(x)$, $T_{i+1}(x)$,...,$T_m(x)$.

In Verteilungen aus m-parametrigen Exponentialfamilien hängt die Verteilung unter der Bedingung, dass die suffizienten Statistiken $T_1(x), \ldots, T_{i-1}(x), T_{i+1}(x), \ldots, T_m(x)$ festgelegte Werte $t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_m$ annehmen, nicht mehr von den zugehörigen Nuisance-Parametern ϑ_1,...,ϑ_{i-1},ϑ_{i+1},...,ϑ_m ab. Diese bedingte Verteilung ist Element der 1-parametrigen Exponentialfamilie, und es existiert somit nach dem Fundamentallemma von Neyman/Pearson ein gleichmäßig bester Test für dieses Problem.

Ein (randomisierter) Signifikanztest Φ ist definiert als eine Abbildung aus dem Stichprobenraum in das Intervall $[0; 1]$. Der Zahlenwert, den Φ annimmt, stellt hierbei die Ablehnwahrscheinlichkeit der Nullhypothese bei einem gegebenen Wert der Testgröße dar (siehe dazu auch z.B. [Witting, 1978]).

$E_\vartheta \Phi^* = P(T \in K)$ bezeichnet den Erwartungswert eines Tests Φ^* unter der Bedingung, dass der Parameterwert ϑ zutrifft. Mit K wird dabei der Ablehnbereich des Tests Φ^* bezeichnet.

Ein nichtrandomisierter Test kann daher als eine Abbildung aus dem Stichprobenraum in die Menge $\{0, 1\}$ definiert werden: Bei Ablehnung der Nullhypothese nimmt der Test Φ den Wert 1 an, ansonsten den Wert 0.

Zunächst stellen wir die allgemeine Form von gleichmäßig besten (unverfälschten) Signifikanztests in einparametrigen Exponentialfamilien gemäß der Neyman/Pearsonschen Testtheorie vor, wie sie z.B. auch in [Lehmann, 1997] oder [Witting, 1978] dargestellt wird. Anschließend zeigen wir, wie im LLTM und ähnlichen Modellen die Verteilung der Testgröße in die Form einer einparametrigen Exponentialfamilie gebracht werden kann.

Punktförmige Nullhypothesen mit zweiseitigen Alternativen

Untersucht werden hier Nullhypothesen der Form

$$H_0 : \quad \vartheta = \vartheta_0$$

gegen die Alternative

$$H_0 : \quad \vartheta \neq \vartheta_0 \quad .$$

Ein gleichmäßig bester unverfälschter Test Φ^* zum Signifikanzniveau α für dieses Testproblem besitzt die folgende Form (zit. nach [Lehmann, 1997], Theorem 4.3.), falls die Verteilung der beobachteten Zufallsgröße X einer einparametrigen Exponentialfamilie entstammt (vgl. dazu auch [Witting, 1978], Kap. 2.8.):

$$\Phi^*(T) = \begin{cases} 1 & falls \quad T(x) < C_1 \vee T(x) > C_2 \\ \gamma_i & falls \quad T(x) = C_i \\ 0 & sonst \quad . \end{cases} \tag{5.3}$$

C_i und γ_i, $i = 1, 2$ werden so bestimmt, dass

$$\begin{aligned} E_{\vartheta_0} \Phi^* &= \alpha \\ E_{\vartheta_0}(\Phi^* T) &= \alpha E_{\vartheta_0} T \end{aligned} \tag{5.4}$$

gilt (vgl. z.B. [Witting, 1978], Kap. 2.8.).

Intervallförmige Nullhypothesen mit zweiseitigen Alternativen

Untersucht werden hier Nullhypothesen der Form

$$H_0: \quad \vartheta_1 \leq \vartheta \leq \vartheta_2$$

gegen die Alternative

$$H_0: \quad \vartheta \notin [\vartheta_1; \vartheta_2] \quad ,$$

wobei $[\vartheta_1; \vartheta_2]$ das abgeschlossene Intervall zwischen den Punkten ϑ_1 und ϑ_2 ist.

In der statistischen Testtheorie von Neyman/Pearson besitzt ein gleichmäßig bester unverfälschter Test zum Signifikanzniveau α für dieses Testproblem die Form

$$\Phi^*(T) = \begin{cases} 1 & falls \quad T(x) < C_1 \vee T(x) > C_2 \\ \gamma_i & falls \quad T(x) = C_i \\ 0 & sonst \quad . \end{cases} \tag{5.5}$$

Dabei werden $C_1, C_2, \gamma_1, \gamma_2$ so bestimmt, dass

$$\begin{aligned} E_{\vartheta_1}\Phi^* &= \alpha \\ E_{\vartheta_2}\Phi^* &= \alpha \end{aligned} \tag{5.6}$$

gilt (vgl. z.B. (vgl. auch [Lehmann, 1997], Kap. 4.2.-4.4.).

Einseitige Nullhypothesen

Untersucht werden hier Nullhypothesen der Form

$$H_0: \quad \vartheta \leq \vartheta_0$$

gegen die Alternative

$$H_A: \quad \vartheta \geq \vartheta_0 \quad .$$

Falls die Verteilung von X einer einparametrigen Exponentialfamilie entstammt, so erhält man einen gleichmäßig besten Test zum Signifikanzniveau α als

$$\Phi^*(T) = \begin{cases} 1 & falls \quad T(x) < C \\ \gamma & falls \quad T(x) = C \\ 0 & sonst \quad . \end{cases} \tag{5.7}$$

Dabei bestimmt man C, γ durch

$$E\Phi^* = \alpha \quad . \tag{5.8}$$

5.3 Die Likelihood des LLTMs als Exponentialfamilie

Im Folgenden werden die Likelihood sowie der gleichmäßig beste Test im Fall zweier Zeitpunkte genauer betrachtet. Hierbei stellen wir dar, unter welchen Bedingungen die Likelihood einer Person bei der Veränderungsmessung als Mitglied einer einparametrigen Exponentialfamilie betrachtet werden kann. Die hierzu durchzuführenden Rechenschritte sind zwar technisch einfach, konnten aber im Zusammenhang mit der Personenfitmessung im LLTM nicht in der Literatur gefunden werden.

Disjunkte Itemparameter

Für die Herleitung nehmen wir zunächst an, dass alle Items an beiden Zeitpunkten dargeboten wurden, und dass nur ein Veränderungsparameter existiert.

Im Fall mit 2 Zeitpunkten besitzt die Likelihood einer Person unter dem LLTM folgende Form:

$$
\begin{aligned}
&P\left(X_v = (x_{v11}, ..., x_{vm1}, ..., x_{vm2}) =\right) \\
&= \prod_{t=1,2} \prod_{i=1,...,m} \frac{exp[x_{vit}(\theta_v - \beta_i + \delta_{gt})]}{1 + exp((\theta_v - \beta_i + \delta_{gt}))} \\
&= C^{-1}(\theta_v, \beta_1, ..., \beta_m, \delta_{g2}) \times exp\left[x_{v\cdot\cdot}\theta_v - \sum_i x_{vi\cdot}\beta_i - \delta_{g2}x_{v\cdot 2}\right] \qquad (5.9) \\
&= C^{-1}(\theta_v, \beta_1, ..., \beta_m, \delta_{g2}) \times exp\left[\sum_i x_{vi\cdot}(\theta_v - \beta_i) - \delta_{g2}x_{v\cdot 2}\right] \quad . \qquad (5.10)
\end{aligned}
$$

Dabei steht δ_{gt} für den Veränderungsparameter im Zeitpunkt $t > 1$ in Gruppe g,
β_i für die Schwierigkeit des Items i,
θ_v für die Fähigkeit einer Person v.
$x_{v\cdot\cdot}$ steht für den Summenscore einer Person v.
$x_{vi\cdot}$ steht für die Anzahl richtiger Lösungen eines Items bei zwei Zeitpunkten, aber einer Person.

Wie man sieht, entstammt die Likelihood für eine Person v unter dem LLTM der m+1-parametrigen Exponentialfamilie von Verteilungen mit (natürlichen) Parametern

$$\theta_v - \beta_1, ..., \theta_v - \beta_m, \delta_{g2} \qquad (5.11)$$

und den dazugehörigen suffizienten Statistiken

$$x_{v1\cdot}, ..., x_{vm\cdot}, x_{v\cdot 2} \quad . \tag{5.12}$$

Wenn wir Hypothesen über δ_{g2} untersuchen wollen, sind alle anderen Parameter nur Nuisance-Parameter, die wir mit Hilfe der suffizienten Statistiken leicht herauspartialisieren können. Wir erhalten somit die bedingte Likelihood

$$\begin{aligned} P(X_v = x_v \mid x_{v1\cdot}, ..., x_{vm\cdot}, \delta_{g2}) = \\ = \frac{\prod_{i=1}^{m} exp(\delta_{g2} x_{vi2})}{\sum_y exp(\delta_{g2} y_{v\cdot 2})} = \\ = \frac{exp(\delta_{g2} x_{v\cdot 2})}{\sum_y exp(\delta_{g2} y_{v\cdot 2})} \quad , \end{aligned} \tag{5.13}$$

die eine Likelihood einer einparametrigen Exponentialfamilie ist. Dabei wird im Nenner des Bruchs über alle Antwortmuster y summiert, die die vorgegebenen Randsummen $(x_{v1\cdot}, ..., x_{vm\cdot})$ besitzen. Für (bestimmte) Hypothesen über δ_{g2} existiert daher nach dem verallgemeinerten Fundamentallemma ein gleichmäßig bester (unverfälschter) Test Φ^* (vgl. dazu [Lehmann, 1997], Kap. 4.2-4.4). Die Wahrscheinlichkeiten $P(X_{v\cdot 2} = l \mid \delta_{g2}, x_{v1\cdot}, ..., x_{vm\cdot})$ lassen sich für den Fall von nur zwei Zeitpunkten folgendermaßen vereinfachen: Für zwei Zeitpunkte kann $x_{vi\cdot}$ die Werte 0 oder 1 oder 2 annehmen. Die n_0 Items mit $x_{vi\cdot} = 0$ tragen nicht zum Summenscore des Zeitpunkts 2 bei, so dass für diesen Summenscore gilt:

$$x_{v\cdot 2} = k + n_2 \leq n_1 + n_2 \quad , \tag{5.14}$$

wobei n_i die Zahl der Items mit Itemscore i über die Zeitpunkte ist. Da die Items mit Itemsummenscore 2 immer zum Summenscore des Zeitpunkts t=2 beitragen, schwankt der Wert von $x_{vi\cdot}$ mit der Zahl k der Items mit $x_{vi1} = 1$, bei denen Item i im zweiten, und nicht im ersten Zeitpunkt gelöst wurde. Dann gibt es genau $\begin{pmatrix} n_1 \\ k \end{pmatrix}$ Antwortmuster, bei denen

$$x_{v\cdot 2} = k + n_2$$

gilt. Zudem besitzt jedes Antwortmuster die gleiche Auftretenswahrscheinlichkeit, da der Nenner für alle Items mit gegebenen Scores eine Konstante ist, und der Zähler nur von der Zahl gelöster Items im zweiten Zeitpunkt abhängt. Somit erhalten wir

$$P(X_{v\cdot 2} = k + n_2 \mid \delta_{g2}, x_{v1\cdot}, ..., x_{vm\cdot}) = \frac{\begin{pmatrix} n_1 \\ k \end{pmatrix} exp[\delta_{g2} k]}{\sum_{l=0}^{n_1} \begin{pmatrix} n_1 \\ l \end{pmatrix} exp[\delta_{g2} l]} \quad , \tag{5.15}$$

der leicht zu berechnen ist. Wie man sieht, entstammt diese Wahrscheinlichkeit einer Binomialverteilung mit Stichprobengröße n_1 und Parameter

$$p_0 = \frac{\exp(\delta_{g2})}{1 + exp(\delta_{g2})} \quad . \tag{5.16}$$

Dies folgt aus

$$\binom{n_1}{k} p^k (1-p)^{n_1-k} \quad = \quad \frac{\binom{n_1}{k} exp(k\delta_{g2})}{\prod_{k=1}^{m}(1 + exp(\delta_{g2}))^m} \quad . \tag{5.17}$$

Im Folgenden bezeichnen wir die Zahl der Items, die im zweiten, nicht aber im ersten Messzeitpunkt beantwortet wurden, mit $x^*_{v\cdot 2}$. Diese Größe verwenden wir als Testgröße zur Konstruktion von gleichmäßig besten (unverfälschten) Signifikanztests.

Nichtdisjunkte Traits mit Interaktionseffekten

Im Folgenden betrachten wir den Fall, dass bei manchen Items die Veränderung mittels mehreren Traits beschrieben werden muss. Wir nehmen o.B.d.A. ein LLTM mit n_1 an beiden Testzeitpunkten ungleich beantworteten Items und 2 Traits δ_1,δ_2 an, wobei bei i_1 Items der Parameter δ_1, bei i_2 Items der Parameter δ_2 und bei $n^* \leq n_1$ Items beide Traits zusammen auftreten. Zur Vereinfachung nehmen wir weiterhin an, dass alle Items durch mindestens einen dieser beiden Traits beeinflusst werden. Als Bezeichnung für einen Interaktionsparameter verwenden wir λ. Die zu den Parametern δ_1 und δ_2 gehörenden suffizienten Statistiken bezeichnen wir mit m_1 resp. m_2.

Im Folgenden bestimmen wir als Beispiel die Verteilung der zu λ gehörenden suffizienten Statistik m_{12} unter der Bedingung, dass die Scoresummen $x_{vi\cdot}$ über die Zeitpunkte sowie m_1 und m_2 gegeben sind. Dabei ist m_1 (resp. m_2 resp. m_{12}) die Zahl der nur im zweiten Messzeitpunkt gelösten Items, in die der Parameter δ_1 (resp. δ_2 resp. λ) in die Berechnung der Lösungswahrscheinlichkeit eingeht.

Als Wahrscheinlichkeit für ein Antwortmuster $\mathbf{X}$ unter der Bedingung vorgegebener Scoresummen $x_{vi\cdot}$ über die Zeitpunkte erhält man dann

$$P(\mathbf{X} \,|\mathbf{x}_{vi\cdot}) = \frac{\exp(m_1\delta_1 + m_2\delta_2 + m_{12}\lambda)}{\prod_{j=1}^{2}(1 + \exp(\delta_j))^{i_j - n^*}(1 + \exp(\delta_1 + \delta_2 + \lambda))} \quad . \tag{5.18}$$

Hieraus ergibt sich die Wahrscheinlichkeit für ein Antwortmuster $\mathbf{X}$, wenn zusätzlich m_1 und m_2 gegeben sind, durch

$$P(\mathbf{X}|\mathbf{x}_{vi\cdot}, m_1, m_2) \quad = \quad \frac{\exp(m_1\delta_1 + m_2\delta_2 + m_{12}\lambda)}{\sum_{l=0}^{n^*}\exp(m_1\delta_1 + m_2\delta_2 + l\lambda)} \tag{5.19}$$

$$= \frac{\exp(m_{12}\lambda)}{\sum_{l=0}^{n^*}\left(\binom{n^*}{l}\exp(l\lambda)\right)} \quad . \tag{5.20}$$

Die Wahrscheinlichkeit, dass genau m_{12} Items mit Interaktionsparameter λ im zweiten Messzeitpunkt gelöst werden, ergibt sich dann durch

$$P(M_{12} = m_{12} \,|\mathbf{x}_{vi\cdot,m_1,m_2}) = \frac{\binom{n^*}{m_{12}}\exp(m_{12}\lambda)}{\sum_{l=0}^{n^*}\left(\binom{n^*}{l}\exp(l\lambda)\right)} \quad . \tag{5.21}$$

Die Verteilung für die Statistik m_{12} ist daher eine Binomialverteilung in

$$p_0 = \frac{\exp(\lambda)}{1+\exp(\lambda)} \tag{5.22}$$

mit Stichprobengröße n^*.

Wie man anhand der hier dargestellten Rechenschritte leicht sieht, resultiert auch bei Tests bezüglich der anderen Veränderungsparameter bei gegebenen Nuisance-Parametern eine Binomialverteilung als bedingte Testgrößenverteilung. Das LLTM mit vollständiger Itemwiederholung kann somit als das Modell charakterisiert werden, bei dem (bedingte) Personenfittests über einen der Veränderungsparameter als Tests bezüglich einer Binomialverteilung durchgeführt werden.

5.4 Gleichmäßig beste Tests für die Personenfitmessung mit Hilfe des LLTMs

Die allgemeine Form eines gleichmäßig besten Signifikanztests in einer einparametrigen Exponentialfamilie wurde im Kapitel 5.2 vorgestellt. In diesem Abschnitt beschäftigen wir uns mit der Frage, wie diese allgemeine Form auf die von uns berechnete Testgrößenverteilung angepasst werden kann.

Im Folgenden bezeichnen wir mit K den Ablehnbereich eines Tests Φ, mit K^a den Rand des Ablehnbereichs und mit K^o das Innere des Ablehnbereichs. Dann gilt für eine diskrete Testgröße TG:

$$E_\vartheta\Phi = \sum_{i\in K^o} Pr(TG = i|\vartheta) + \sum_{c_i\in K^a} Pr(TG = c_i|\vartheta)\gamma_i \quad . \tag{5.23}$$

Im Fall des Binomialtests ergibt sich $Pr(TG = i|\vartheta)$ stets gemäß (5.16) und (5.15) aus einer Binomialverteilung.

$E\Phi T$ berechnet man durch

$$E_\vartheta \Phi = \sum_{t \in K^o} tPr(TG = t|\vartheta) + \sum_{t=c_i \in K^a} Pr(TG = t|\vartheta)\gamma_i \quad . \tag{5.24}$$

Die genaue Spezifizierung der besten Tests ergibt sich nun durch die Form des Ablehnbereichs.

Zweiseitige Nullhypothesen beim Binomialtest

Die suffiziente Statistik $T(x)$ aus Kap. 5.2 kann mit $x^*_{v\cdot 2}$ gleichgesetzt werden. Mit $\vartheta = \delta_0$ aus der Nullhypothese und (5.23) sowie (5.24) kann der Ablehnbereich K bestimmt werden. Dieser besteht aus den zwei Intervallen $[0; c_1]$ und $[c_2; n_1]$. Durch Einsetzen in die Gleichungen (5.3) erhält man die Bedingungen, die für den hier dargestellten Fall zur Bestimmung gleichmäßig bester unverfälschter Tests benötigt werden.

Intervallförmige Nullhypothesen

Im Fall der Gleichungen (5.6) besitzt die Nullhypothese die Form $\delta_1 \geq \delta_{g2} \leq \delta_2$. Hier erhält man p_0, indem man $\delta_{g2} = \delta_1$ (resp. $\delta_{g2} = \delta_2$) in Gleichung (5.16) einsetzt.

Für δ_1 erhält man:

$$p_1 = \frac{\exp(\delta_1)}{1 + exp(\delta_1)} \quad , \tag{5.25}$$

für δ_2:

$$p_2 = \frac{\exp(\delta_2)}{1 + exp(\delta_2)} \quad . \tag{5.26}$$

Die Ablehnbereiche ergeben sich durch Einsetzen von δ_1 resp. δ_2 in (5.23) und (5.24), sowie durch Anwendung der Bedingungen aus (5.6). Der Ablehnbereich besteht aus zwei Intervallen $[0; c_1]$ und $[c_2; n_1]$.

Einseitige Tests

Für den Fall des einseitigen Tests (5.8) nehmen wir zunächst an, dass die Nullhypothese $\vartheta \leq \vartheta_0$ gegen die Alternative $\vartheta > \vartheta_0$ getestet wird. Hierbei setzt man δ_{g2} anstelle von ϑ in (5.23) bzw. (5.24) ein. Der Ablehnbereich besteht dann aus einem Intervall $[c; n_1]$.

Die Formeln für das umgekehrte Hypothesenpaar können hieraus leicht abgeleitet werden. Man erhält p_0, indem man den Rand ϑ_0 der Nullhypothese anstelle von δ_{g2} in Gleichung

(5.16) einsetzt.

Zur Bestimmung von γ_i und C_i

Die Bestimmung der C_i, γ_i kann nach verschiedenen Methoden erfolgen. Am einfachsten ist die Lage bei den einseitigen Tests, für die Bedingung (5.8) erfüllt sein muss. Hier existiert bei gegebenem α nur ein C, für das die Bedingung überhaupt erfüllt sein kann, für dieses C löst man nach γ auf.

Für die zweiseitigen Tests (5.3) und (5.6) ist die Situation nicht so eindeutig. Am naheliegendsten ist hierbei die Methode, Paare (C_1, C_2) vorzugeben, für die eine Teilbedingung erfüllt ist, und dann bei vorgegebenem α nach γ_1, γ_2 aufzulösen. Wegen der durch das verallgemeinerte Fundamentallemma garantierten Eindeutigkeit der gleichmäßig besten Tests (vgl. z.B. [Witting, 1978]) existiert für genau ein Paar (C_1, C_2) eine Lösung mit $0 \leq \gamma_1, \gamma_2 \leq 1$. Diese Vorgehensweise kann jedoch recht aufwändig sein, zumal bei großem n_1. [Witting, 1978], S. 102 gibt daher rechnerische Vereinfachungen für gleichmäßig beste Tests bei Binomialverteilungen an, auf die wir hier nicht weiter eingehen.

[Klauer, 1991b] betrachtet die Größe

$$U(X) = X^*_{v\cdot 2} + Z \quad , \tag{5.27}$$

wobei Z gleichverteilt auf dem Intervall $[0, 1[$ ist. U(X) besitzt eine Dichte

$$f(u) = P(X^*_{v\cdot 2} = x^*_{v\cdot 2}) \quad , \tag{5.28}$$

falls $x^*_{v\cdot 2} \leq u < x^*_{v\cdot 2} + 1$ gilt.

Da U definiert ist für alle $0 \leq U < m + 1$ und somit nicht mehr diskret ist, kann obiger Test in einen Test umgeformt werden, der auf U beruht und ohne die Randomisierungsparameter γ_i auskommt. Ein solcher Test verwendet statt einer Schranke C_i mit Randomisierungsparameter γ_i aus obigem Test die Grenze $\epsilon_i = C_i + \gamma_i$. Falls man jetzt U statt $X_{v\cdot 2}$, sowie ϵ statt C und γ in obiges Gleichungssystem einsetzt, kann man mit einfachen numerischen Methoden eine eindeutige Lösung bestimmen (vgl. [Klauer, 1991b]).

5.5 Anmerkungen zu den vorgeschlagenen Verfahren

Diskutiert werden müssen in diesem Zusammenhang drei Punkte: die Angemessenheit eines randomisierten Testverfahrens, die Verwendung der intervallförmigen Nullhypothese und die Verwendung von Interaktionshypothesen wie in [Klauer, 1995].

Angemessenheit randomisierter Tests

[Klauer, 1995] diskutiert die Angemessenheit randomisierter Tests in der psychologischen Forschungspraxis. Er beschreibt, dass für einen Forscher eine Testentscheidung aufgrund eines Zufallsexperiments nicht einsichtig ist, da durch eine Zufallsentscheidung Personen trotz gleicher Symptome in unterschiedliche Klassen eingeordnet werden. Für diesen Fall empfiehlt Klauer die Verwendung nichtrandomisierter Tests. Diese lassen sich aus den beschriebenen randomisierten Testverfahren herleiten, indem man $\gamma_i = 1$ setzt. Der so erzeugte Test ist konservativer als der zugehörige randomisierte Test, kann jedoch als gleichmäßig bester nichtrandomisierter Test angesehen werden.

Die Verwendung intervallförmiger Nullhypothesen

Intervallförmige Nullhypothesen sind – nach Kenntnis des Autors – derzeit in der Psychologie nicht gebräuchlich. Gleichwohl bieten intervallförmige Nullhypothesen einige Vorteile gegenüber normalen zweiseitigen Signifikanztests und könnten dabei helfen, das z.B. in [Sedlmeier, 1996] geäußerte Unbehagen an der Verwendung von Signifikanztests als psychologischer Forschungsmethode zu vermindern.

Intervallförmige Nullhypothesen ermöglichen es nämlich, den Gedanken der „praktischen Relevanz“ eines signifikanten Ergebnisses in die Untersuchungsmethode einfließen zu lassen. Dazu nehmen wir eine Nullhypothese $\vartheta = \vartheta_0$ an. Innerhalb eines Intervalls

$$[\vartheta_0 - \sigma; \vartheta_0 + \sigma]$$

liegen Abweichungen von der Nullhypothese vor, die man aber aus nichtstatistischen Gründen als nicht relevant einstuft. Die Verwendung einer Nullhypothese

$$\vartheta \in [\vartheta_0 - \sigma; \vartheta_0 + \sigma]$$

vereinfacht die Entscheidung, ob man das Ergebnis eines Experiments als relevante Abweichung von der (ursprünglichen) Nullhypothese beurteilen soll, oder nicht. Intervallförmige Nullhypothesen können somit als Ergänzung zur Verwendung von Effektgrößen o.ä. Verfahren dienen (vgl. dazu auch [Sedlmeier, 1996]).

Verwendung der optimalen Personenfittests im LLRA

In diesem Abschnitt zeigen wir auf, dass sich die in diesem Kapitel aus dem LLTM abgeleiteten Personenfittests auch im Rahmen des LLRAs nutzen lassen. Wie in Kapitel 2.2 dargestellt, verzichtet das LLRA auf die Definition getrennter Item- und Fähigkeitsparameter. Stattdessen wird ein Parameter θ_{vi} verwendet, der die Schwierigkeit des Items i bei Person v beschreibt.

Die Likelihood des LLRAs ist Element der Exponentialfamilie in θ_{vi} und δ_{gt}. Dies folgt aus der Grundgleichung des LLRAs (2.3). Suffiziente Statistik für die Parameter θ_{vi} ist dabei $X_{vi\cdot}$.

Durch Berechnen der bedingten Likelihood unter der Bedingung

$$X_{vi\cdot} = x_{vi\cdot}$$

erhält man, wie man leicht sieht, die gleiche bedingte Likelihood wie in (5.13). Somit können die in diesem Kapitel eingeführten Testverfahren auch für das LLRA verwendet werden.

Kapitel 6

Tests für allgemeinere Untersuchungssituationen

In diesem Abschnitt wird untersucht, wie die bisher vorgestellten Methoden verallgemeinert werden können.

6.1 Ein allgemeines loglineares Modell zur Veränderungsmessung

Wir werden im Folgenden eine vereinfachte Version des allgemeinen Rasch-Modells zur Veränderungsmessung verwenden, wie es in [Meiser, 1996] vorgestellt wurde. Dieses Modell erlaubt in seiner Originalform

1. polytome ordinal skalierte Items,
2. mehrere Zeitpunkte,
3. mehrere Traits, sowie
4. unterschiedliche Veränderungen für jede Stufe eines Items.

Wir betrachten zunächst eine vereinfachte Version, bei der nur ein Trait, zwei Zeitpunkte und eine konstante Veränderung pro Person für alle Items angenommen werden.

Bevor wir diese Verallgemeinerung einführen, müssen wir einige für polytome Modelle spezifische Modellvoraussetzungen erläutern: Polytome Item-Response-Modelle für ordinale Items gehen von der Existenz sogenannter Schwellenparameter aus. Um dieses Konzept zu erläutern, nehmen wir ein $J+1$-stufiges Item mit den Stufenwerten $0, 1, \ldots, J$ an.

Modelliert wird die Wahrscheinlichkeit, dass eine Person sich für eine Stufe j entscheidet, unter der Bedingung, dass sie sich für Stufe $j-1$ oder Stufe j entscheidet:

$$P(X_{vi} = j | X_{vi} = j \wedge X_{vi} = j-1) = \frac{exp(\theta_{vi} - \tau_{ij})}{1 + exp(\theta_v - \tau_{ij})} \quad . \tag{6.1}$$

Dabei ist θ_v die Fähigkeit einer Person v, τ_{ij} der Schwellenwert für Stufe j des Items i (vgl. dazu [Rost and v. Davier, 1995]). Die (bezüglich der möglichen Antwortstufen) unbedingte Wahrscheinlichkeit für die Wahl der Stufe j des Items i ergibt sich aus

$$P(X_{vi} = j \mid \theta_v,) = \frac{exp(j\theta_v - \sum_{h=1}^{j} \tau_{ih})}{1 + \sum_{h=1}^{J} exp(h\theta_v - \sum_{h=1}^{j} \tau_{ih})} \tag{6.2}$$

(vgl. dazu ebenfalls [Rost and v. Davier, 1995]). Als daraus aufgebautes Veränderungsmodell verwenden wir (vgl. [Fischer and Ponocny, 1994]) folgende einfache Version des LPCM:

$$\begin{aligned} P(\mathbf{X} = \mathbf{x} | \theta_v, \tau_{111}, \ldots, \tau_{2mJ}) = \\ = \frac{\exp\left(x_{v\cdot 1}\theta_v - \sum_{i=1}^{m}\sum_{h=1}^{x_{vi1}} \tau_{1ih} + x_{v\cdot 2}\theta_v - \sum_{i=1}^{m}\sum_{h=1}^{x_{vi2}} \tau_{1ih} + \sum_{i=1}^{m} x_{2vi}\delta_{2g}\right)}{\prod_{t=1}^{2}\prod_{i=1}^{m}\sum_{j=0}^{J}\exp\left(j\theta_v - \sum_{h=1}^{j}\tau_{ij}\right)} \quad . \end{aligned} \tag{6.3}$$

Um einen optimalen Signifikanztest für polytome Items zu entwickeln, führen wir die Indikatorvariable

$$C_{vith} = \begin{cases} 1 & : \quad X_{vit} \leq h \\ 0 & : \quad X_{vit} > h \end{cases} \tag{6.4}$$

ein. Die zu C_{vith} gehörende Realisation bezeichnen wir mit c_{vith}.

Mit Hilfe dieser Indikatorvariablen erhalten wir eine Verteilung für $\mathbf{X}$, die Element der Exponentialfamilie mit den kanonischen Parametern $\theta_v - \tau_{1ih}$ $(1 \leq i \leq m, 1 \leq h \leq J)$ und δ_{2g} sowie zugehörigen suffizienten Statistiken $(C_{vi1h} + C_{vi2h})$, resp. $\sum_{i=1}^{m} X_{vi2h}$ ist:

$$P(\mathbf{X}=\mathbf{x}|\theta_v,\tau_{111},\ldots,\tau_{2mJ}) =$$
$$= \frac{\exp\left[\sum_{i=1}^{m}\sum_{h=1}^{J}((c_{vi1h}+c_{vi2h})(\theta_v-\tau_{1ih}))+\sum_{i=1}^{m}x_{2vi}\delta_{2g}\right]}{\prod_{t=1}^{2}\prod_{i=1}^{m}\sum_{j=0}^{J}\exp\left(j\theta_v-\sum_{h=1}^{j}\tau_{ij}\right)} \quad . \tag{6.5}$$

Dabei sind die $\theta_v - \tau_{1ih}$ Nuisance-Parameter, falls wir an der Veränderung δ_{2g} interessiert sind. Ebenso wie in Kapitel 5.3 können wir für Hypothesen über δ_{2g} bedingte Tests bei vorgegebenen Werten für $(c_{vi1h}+c_{vi2h})$ konstruieren. Als bedingte Wahrscheinlichkeit für $\mathbf{X}=\mathbf{x}$ unter der Bedingung fest vorgegebener suffizienter Statistiken für die Nuisance-Parameter erhalten wir dann

$$P(\mathbf{X}=\mathbf{x}|\sum_t C_{v1t1}=c_{v111}+c_{v121},\ldots,\sum_t C_{vmtJ}=c_{vm1J}+c_{vm2J}) =$$
$$= \frac{\exp\left[\sum_{i=1}^{m}\sum_{h=1}^{J}((c_{vi1h}+c_{vi2h})(\theta_v-\tau_{1ih}))+\sum_{i=1}^{m}x_{2vi}\delta_{2g}\right]}{\sum\exp\left[\sum_{i=1}^{m}\sum_{h=1}^{J}((c_{vi1h}+c_{vi2h})(\theta_v-\tau_{1ih}))+\sum_{i=1}^{m}x_{2vi}\delta_{2g}\right]} \tag{6.6}$$
$$= \frac{\exp\left[\sum_{i=1}^{m}x_{2vi}\delta_{2g}\right]}{\sum\exp\left[\sum_{i=1}^{m}x_{2vi}\delta_{2g}\right]} \quad . \tag{6.7}$$

Die Summe im Nenner der obigen Brüche wird dabei immer über alle Antwortmuster mit vorgegebenen Werten für die suffizienten Statistiken der Nuisamce-Parameter berechnet. Mit einer zu Kapitel 5.3 parallelen Argumentationskette kommen wir zu einer Binomialverteilung für $\sum_i X_{2vi} = \sum \mathbf{x}_{2vi}$ unter der Bedingung, dass

$$\sum_t C_{v1t1}=c_{v111}+c_{v121},\ldots,\sum_t C_{vmtJ}=c_{vm1J}+c_{vm2J} \tag{6.8}$$

gilt:

$$P\left(\sum\nolimits_i X_{2vi}=\sum\mathbf{x}_{2vi}|\sum_t c_{v1t1},\ldots,\sum_t c_{vmtJ}\right) = \frac{\binom{n_1}{k}exp[\delta_{g2}k]}{\sum_{l=0}^{n_1}\binom{n_1}{l}exp[\delta_{g2}l]} \quad . \tag{6.9}$$

Hierbei ist n_1 die Zahl der Paare (C_{vi1h}, C_{vi2h}), für die

$$C_{vi1h}+C_{vi2h}=1$$

gilt, sowie k die Zahl der Paare (C_{vi1h}, C_{vi2h}) mit $C_{vi2h} = 1$ und $C_{vi1h} = 0$.

Analog zu Kapitel 5.3 stellt (6.9) eine Binomialverteilung für k mit

$$p = \frac{exp(\delta_{g2})}{1 + exp(\delta_{g2})} \tag{6.10}$$

und Stichprobenumfang n_1 dar.

6.2 Tests für Untersuchungsdesigns mit mehr als zwei Zeitpunkten

Dieser Abschnitt beschreibt, wie die dargestellten Testmethoden auf mehr als zwei Messzeitpunkte verallgemeinert werden können. Für mehr als zwei Zeitpunkte können nur unter sehr restriktiven Bedingungen optimale Signifikanztests hergeleitet werden: Da für jeden Zeitpunkt ein Veränderungsparameter existiert, ist die Likelihood bei mehr als zwei Zeitpunkten echt mehrdimensional. Statistisch optimale Signifikanztests existieren im Rahmen der Neyman-Pearsonschen Signifikanztesttheorie nicht für solche echt mehrdimensionalen Testprobleme (vgl. [Lehmann, 1997], [Witting, 1985]). Daher werden wir in diesem Abschnitt exakte, aber nichtoptimale Signifikanzests für die Untersuchung dieses Problems vorstellen. Weiterhin werden wir Tests für Spezialfälle (wie z.B. monotoner Trend des Veränderungsparameters) vorstellen.

Die bedingte Likelihood im Fall von mehr als zwei Zeitpunkten

Zunächst verallgemeinern wir die Likelihood (5.10). Bei $T > 2$ Zeitpunkten erhalten wir:

$$\begin{aligned}
&P\left(X_v = (x_{v11}, ..., x_{vm1}, ..., x_{vmT}) =\right) \\
&= \prod_{t=1}^{T} \prod_{i=1,...,m} \frac{exp[x_{vit}(\theta_v - \beta_i + \delta_{gt})]}{1 + exp((\theta_v - \beta_i + \delta_{gt}))} \\
&= C^{-1}(\theta_v, \beta_1, ..., \beta_m, \delta_{g2}, \cdot, \delta_{gT}) \times exp\left[x_{v\cdot\cdot}\theta_v - \sum_i x_{vi\cdot}\beta_i - \sum_{t=2}^{T} \delta_{gt} x_{v\cdot t}\right] \qquad (6.11) \\
&= C^{-1}(\theta_v, \beta_1, ..., \beta_m, \delta_{g2}\cdot, \delta_{gT}) \times exp\left[\sum_i x_{vi\cdot}(\theta_v - \beta_i) - \sum_{t=2}^{T} \delta_{gt} x_{v\cdot t}\right] \quad . \qquad (6.12)
\end{aligned}$$

Diese Likelihood ist Element der Exponentialfamilie in

$$\theta_v - \beta_1, \theta_v - \beta_2, ..., \theta_v - \beta_m; \delta_{g2}, ..., \delta_{gT}$$

mit suffizienten Statistiken

$$X_{v1\cdot}, ..., X_{vm\cdot} \text{ und } X_{\cdot\cdot 2}, ..., X_{\cdot\cdot T} \quad .$$

Wenn wir die bedingte Likelihood unter

$$X_{v1\cdot} = x_{v1\cdot}, ..., X_{vm\cdot = x_{vm\cdot}}$$

berechnen, erhalten wir ein Testproblem für eine 2^T-Kontingenztafel, in dem nur noch die interessierenden Parameter $\delta_{g2}, ..., \delta_{gT}$ enthalten sind:

$$\begin{aligned} & P(X_v = x_v \mid x_{v1\cdot}, ..., x_{vm\cdot}, \delta_{g2}, \ldots, \delta_{gT}) = \\ = & \frac{\prod_{i=1}^{m} \prod_{t=2}^{T} exp(\delta_{gt} x_{vit})}{\prod_{i=1}^{m} \prod_t (1 + \exp(\delta_{gt}))} = \\ = & \frac{exp(\sum_t \delta_{gt} x_{v\cdot t})}{\prod_{i=1}^{m} \prod_t [1 + \exp(\delta_{gt})]} \quad . \end{aligned} \tag{6.13}$$

Zunächst betrachten wir das echt mehrparametrige Testproblem bezüglich der Parameter $\delta_{g2}, ..., \delta_{gT}$. Nullhypothese ist dann:

$$H_0 : \delta_{g2} = \Delta^0_{g2}, ..., \delta_{gT} = \Delta^0_{gT} \quad .$$

Dies wird gegen die Alternative „Für mindestens ein δ_{gt} gilt: $\delta_{gt} \neq \Delta^0_{gt}$“ getestet. Hier liegt eine echt mehrparametrige Nullhypothese vor. Eine Aufspaltung in einen interessierenden Parameter und mehrere Nuisance-Parameter ist nicht möglich. Daher kann hier auch kein nach dem verallgemeinerten Fundamentallemma gleichmäßig bester Test konstruiert werden (vgl. [Lehmann, 1997], [Witting, 1978]).

Für ein solches Testproblem lässt sich ein exakter Likelihood-Ratio-Test verwenden. Die Daten lassen sich in einer $\left(2^T\right) \times m$-Kontingenztafel anordnen. Jede Zelle dieser $\left(2^T\right) \times m$-Kontingenztafel entspricht einem Antwortvektor $X_{vi1}, ..., X_{viT}$ einer Person v bei einem Item i. Für jede Zelle dieser Kontingenztafel erhält man die Auftretenswahrscheinlichkeit durch

$$Pr(X_{vi1} = x_{vi1}, ..., X_{viT} = x_{viT}) = \frac{\exp(\sum_t \delta_{gt} x_{vit})}{\prod_t (1 + \exp(\delta_{gt})} \times Pr(X_{v1\cdot} = x_{v1\cdot}, ..., X_{vm\cdot} = x_{vm\cdot}) \quad . \tag{6.14}$$

Mit r_{vi} ist hierbei die Zahl der richtigen Antworten einer Person v bei Item i über alle Zeitpunkte hinweg bezeichnet. Die Testgröße des Likelihood-Ratio-Tests erhält man durch

$$G^2 = -2m\left[\ln(P(X_v = x_v \mid x_{v1\cdot}, ..., x_{vm\cdot}, \Delta_{g2}, \cdot, \Delta_{gT})) - \ln(P(X_v = x_v \mid x_{v1\cdot}, ..., x_{vm\cdot})\right] \quad . \tag{6.15}$$

Hierbei ist $P(X_v = x_v \mid x_{v1\cdot}, ..., x_{vm\cdot})$ die Likelihood des saturierten Modells, für das gilt:

$$Pr(X_{vi1} = x_{vi1}, ..., X_{viT} = x_{viT}) = Pr(\sum_t X_{vit} = r_{vi}) = \frac{m_{x_{vi}}}{m} \quad . \tag{6.16}$$

$m_{x_{vi}}$ ist die Zahl der Items (bzw. im polytomen Fall: Schwellen), bei denen das Antwortmuster $x_{vi} = (x_{vi1}, \ldots, x_{viT})$ vorkommt. m ist die Gesamtzahl der Items (bzw. Schwellen) im Fragebogen. Die approximative χ^2-Verteilung der Testgröße folgt daraus, dass die Vektoren

$$(X_{vi1}, ..., X_{viT})$$

unter der Bedingung gegebener Randsumme $\sum_t X_{vit} = r_{vi}$ unabhängig voneinander sind. Daher ist die Verteilung der gesamten Kontingenztafel, also des Vektors

$$(X_{v11}, ..., X_{v1T}, \ldots, X_{vi1}, ..., X_{viT}, \ldots X_{vm1}, ..., X_{vmT})$$

das Produkt mehrerer unabhängiger Multinomialverteilungen. Die Testgröße G^2 ist folglich approximativ χ^2-verteilt mit $m\left(2^T - 1\right)$ Freiheitsgraden.

Tests auf linearen Trend

An dieser Stelle wird ein Test auf linearen Trend der Veränderungsparameter δ_{gt} bezüglich des Messzeitpunkts t vorgestellt. In diesem Spezialfall kann ein statistisch optimaler Signifikanztest hergeleitet werden. Ausgangspunkt hierbei ist

$$\begin{aligned} P(X_v = x_v \mid x_{v1\cdot}, ..., x_{vm\cdot}, \delta_{g2}, \ldots, \delta_{gT}) = \\ = \quad & \frac{exp(\sum_t \delta_{gt} x_{v\cdot t})}{\prod_{i=1}^m \prod_t \left[1 + \exp\left(\delta_{gt}\right)\right]} \\ = \quad & \frac{exp(\sum_t \delta t x_{v\cdot t})}{\prod_{i=1}^m \prod_t \left[1 + \exp\left(\delta_{gt}\right)\right]} \quad . \end{aligned} \tag{6.17}$$

Hierbei wird angenommen, dass δ der Steigungsparameter einer linearen Regression

$$\delta_{gt} = \delta t + \epsilon$$

ist. Wie man sieht, stellt Gleichung (6.17) eine einparametrige Exponentialfamilie im Parameter δ und der suffizienten Statistik $\sum_t t x_{v\cdot t}$ dar. Daher kann auch in diesem Fall ein gleichmäßig bester (unverfälschter) Test für die Nullhypothese

$$\delta = \Delta_0$$

oder für andere der in Kapitel 5 aufgeführten Nullhypothesentypen konstruiert werden.

6.3 Multivariate Veränderungen

In diesem Abschnitt werden Personenfittests für den folgenden Fall beschrieben: Es sind zwei Zeitpunkte vorgegeben, es müssen aber mindestens zwei hinsichtlich der Veränderung konstante Itemgruppen angenommen werden.

Hierzu setzen wir ein gegenüber dem allgemeinen LLTM (2.2) vereinfachtes Modell voraus, in dessen Parametrisierung keine Interaktionen vorkommen: Wir gehen von der Existenz mehrerer Veränderungsparameter $\delta_{12}, \ldots, \delta_{h2}$ aus. Dabei nehmen wir an, dass jedem Item i eindeutig ein Veränderungsparameter δ_{k2} zugeordnet werden kann. Mit I_k bezeichnen wir die Menge der Items, die einem Veränderungsparameter δ_{k2} zugeordnet sind. Weiterhin nehmen wir an, dass für jeden Parameter δ_{i2} eine suffiziente Statistik T_{i2} existiert. Es wird vorausgesetzt, dass die suffizienten Statistiken $T_{12}, \ldots, T_{k2}$ linear unabhängige Funktionen sind (vgl. z.B. [Witting, 1978]). Schließlich setzen wir voraus, dass die Auftretenswahrscheinlichkeit

$$P\left[X_v = (x_{v11}, \ldots x_{vm1}, x_{v12}, \ldots, x_{vm2})\right]$$

Mitglied einer k+m-parametrigen Exponentialfamilie ist:

$$\begin{aligned} P\left[X_v = (x_{v11}, \ldots x_{vm1}, x_{v12}, \ldots, x_{vm2})\right] = & \\ = \quad & \frac{\exp\left[\sum_t \sum_i x_{vit}\left(\theta_v - \beta_i + \delta_{kt}\right)\right]}{1 + \exp\left[\sum_t \sum_i x_{vit}\left(\theta_v - \beta_i + \delta_{kt}\right)\right]} \qquad (6.18) \\ = \quad & \frac{\exp\left[\sum_i \left[\sum_t x_{vit}\left(\theta_v - \beta_i\right)\right] + \sum_k \sum_{i \in I_k}\left(x_{vi2}\delta_{k2}\right)\right]}{C(\theta_v, \delta_1, \ldots, \delta_k, \beta_1, \ldots, \beta_m)} \; . \qquad (6.19) \end{aligned}$$

Hierbei ist

$$T_{12} = \sum_{i \in I_1} (x_{vi2}) \tag{6.20}$$

$$\vdots \quad \vdots$$

$$T_{1k} = \sum_{i \in I_k} (x_{vi2}) \tag{6.21}$$

$C(\theta_v, \delta_1, \ldots, \delta_k, \beta_1, \ldots, \beta_m)$ ist eine Konstante im Nenner, die bei der Konstruktion bedingter Tests unter der Nullhypothese verschwindet. Jedem Item i kann also eine Veränderung δ_{k2} zugeordnet werden. Auf diese Weise kann die lineare Unabhängigkeit der suffizienten Statistiken sichergestellt werden. Dies führt i.d.R. zu einem echt mehrparametrigen Testproblem, das nicht auf ein Problem der Art „k-1 Nuisance-Parameter, 1 interessierender Parameter“ reduziert werden kann. Gleichmäßig beste (unverfälschte) Tests existieren dann nicht (vgl. z.B. [Witting, 1978], Kap. 4). Im Folgenden wird ein wichtiger Fall beschrieben, bei dem die Reduktion auf ein eindimensionales Testproblem möglich ist. Für andere Fälle werden (exakte) Likelihood-Quotiententests vorgeschlagen.

6.3.1 Einfache Kontrasthypothesen

Dieser Abschnitt beschäftigt sich mit Hypothesen der Form

$$\delta_{j_1 2} = \delta_{j_2 2} = \ldots \delta_{j_s 2} = \delta^* \ ,$$

die als Kontrasthypothesen formuliert werden können. δ^* ist dabei unbekannt. Gegenhypothese ist dabei: Für mindestens einen Veränderungsparameter δ_{k2}, $k \in \{j_1, \ldots, j_s\}$, gilt:

$$\delta_{k2} \neq \delta^* \ .$$

Auch bei dieser Art von Testproblemen können gleichmäßig beste (unverfälschte) Tests abgeleitet werden. Dazu verwenden wir s als die Zahl der miteinander zu vergleichenden Veränderungsparameter. Die Wahrscheinlichkeit eines Antwortmusters ergibt sich als Mitglied der Exponentialfamilie mit den folgenden Parametern und dazugehörigen suffizienten Statistiken:

- Parameter $\theta_v - \beta_i$ und suffiziente Statistik x_{vi+} (Für $i = 1, \ldots, m$).
- Parameter δ_{j2} und zugehörige suffiziente Statistik $\sum_{i \in I_j} x_{vi2}$ für $j \notin \{j_1, \ldots j_s\}$.
- Parameter δ^* und zugehörige suffiziente Statistik $\sum_i \sum_{i \in I_j} x_{vi2}$ für $j \in \{j_1, \ldots j_s\}$.

Die Wahrscheinlichkeit für ein Antwortmuster X_v ergibt sich bei Gültigkeit der Nullhypothese aus (6.19), wenn man für die Parameter $\delta_{j_1 2}, \ldots, \delta_{j_s 2}$ den (unbekannten) Wert δ^* einsetzt:

$$\begin{aligned} P\left[X_v = (x_{v11}, \ldots x_{vm1}, x_{v12}, \ldots, x_{vm2})\right] = \\ = \frac{\exp\left[\sum_i \left[x_{vi+}\left(\theta_v - \beta_i\right)\right] + \sum_{k \notin \{j_1,\ldots,j_s\}} t_{k2}\delta_{k2} + \sum_{k \in \{j_1,\ldots,j_s\}} t_{k2}\delta_{k2}\right]}{C(\theta_v, \delta_1, \ldots, \delta_k, \beta_1, \ldots, \beta_m)} \qquad (6.22) \\ = \frac{\exp\left[\sum_i [x_{vi+}(\theta_v - \beta_i)] + \sum\limits_{k \notin \{j_1,\ldots,j_s\}} t_{k2}\delta_{k2} + \sum\limits_{j_i} t_{j_i 2}\left(\delta_{j_i 2} - \delta_{j_s 2}\right) + \sum\limits_{j_i} t_{j_i 2}\delta_{j_s 2}\right]}{C(\theta_v, \delta_1, \ldots, \delta_{k2}, \beta_1, \ldots, \beta_m)} \; . \qquad (6.23) \end{aligned}$$

Die Statistik $\sum_{k \in \{j_1,\ldots j_s\}} t_{k2}$ enthält die gesamte Information über den unbekannten Parameter δ^*. Das weitere Vorgehen ähnelt der Herleitung des exakten Tests von Fischer (vgl. z.B. [Witting, 1978]). In (6.23) sind $\delta_{j_1 2} - \delta_{j_s 2}, \ldots, \delta_{j_{s-1} 2} - \delta_{j_s 2}$ die interessierenden Parameter, δ_{j_s} ein weiterer Nuisanceparameter. Bei Gültigkeit der Nullhypothese und gegebenen Werten für x_{vi+} nehmen letztere Parameter den Wert 0 an. Als bedingte Verteilung des Vektors

$$\mathbf{T} = \left(T_{j_1 2}, \ldots, T_{j_{s-1} 2}\right)$$

unter der Nullhypothese erhält man mit $N_{j_i} = |I_{j_i}|$ als Zahl der Items für Veränderungsparameter δ_{j_i}, $m = \sum N_{j_i}$, sowie $T^* = \sum T_{j_i 2}$ und $j_i \in \{j_1, \ldots, j_s\}$:

$$\begin{aligned} P\left[\mathbf{T} = \left(t_{j_1 2}, \ldots, t_{j_{s-1} 2}\right) | x_{v1+}, \ldots, x_{vm\cdot}, t_{k_1 2}, \ldots, t_{k_{m-s} 2}, \sum_{k \notin \{j_1,\ldots,j_s\}} t_{j_i 2}\right] = \\ = \frac{\binom{t^*}{t_{j_1 2}, \ldots, t_{j_s 2}} \binom{m - t^*}{N_{j_1} - t_{j_1 2}, \ldots, N_{j_s} t_{j_s 2}}}{\binom{m}{N_{j_1}, \ldots, N_{j_s}}} \; . \qquad (6.24) \end{aligned}$$

Hierbei umfasst der Indexvektor $(k_1, \ldots, k_{m-s})$ alle Veränderungsparameter, die nicht von der Nullhypothese betroffen sind. Im Nenner dieses Bruchs wird dabei über alle $\binom{\sum_{j_i < j_s} T_{j_i 2}}{T_{j_1 2}, \ldots, T_{j_{s-1} 2}}$ summiert, die bei gegebenen suffizienten Statistiken auftreten können. Dieser Ansatz führt, wie man leicht sieht,

- im Fall s=2 (d.h. nur zwei verschiedener Veränderungsparameter) zum exakten Test von Fisher (vgl. zur Herleitung des exakten Tests von Fisher auch [Witting, 1978]). Dieser Test ist gleichmäßig bester unverfälschter Test für obiges Testproblem. Hierbei ist T_1 die Testgröße, die hypergeometrisch verteilt ist in m und N_1.

- im Fall $s > 2$§ zum exakten Test von [Freeman and Halton, 1951] für $2 \times s$-Kontingenztafeln. Dieser Test verwendet (im Fall s=2) die gleiche Testgrößenverteilung wie der exakte Test von Fisher, bestimmt jedoch den Ablehnbereich derartig, dass auch eine Verallgemeinerung auf den Fall $s > 2$ möglich ist. Der Ablehnbereich dieses Tests zum Signifikanzniveau α wird durch die $100 \times \alpha\%$ Antwortmuster mit der niedrigsten Likelihood (6.24) bestimmt. Der Freeman/Halton-Test besitzt nicht die Optimalitätseigenschaften des Fisher-Tests.

6.3.2 Exakte Likelihood-Ratiotests für allgemeine Hypothesen

Für allgemeine Hypothesen der Form

$$H_0 : \delta_{j_1 2} = \delta_{j_2 2} = \ldots \delta_{j_h 2} = c_1, \ldots, \delta_{j_r 2} = \ldots = \delta_{j_s 2} = c_z \tag{6.25}$$

kann kein gleichmäßig bester Test angegeben werden, da das Testproblem nicht auf einen interessierenden Parameter reduziert werden kann. Ausgehend von (6.19) lassen sich hier bedingte Tests mit der Likelihood als Testgröße konstruieren. Nuisance-Parameter sind hierbei:

- Alle Veränderungsparameter $\delta_{k_1}, \ldots, \delta_{k_h}$, die nicht in (6.25) aufgeführt sind.
- Alle Fähigkeitsparameter θ_v
- Alle Schwierigkeitsparameter β_i

Diese Parameter tauchen nicht in der Nullhypothese auf. Durch die Verwendung der bedingten Likelihood (unter der Bedingung vorgegebener suffizienter Statistiken für die Nuisanceparameter) können die Nuisance-Parameter aus der Likelihood herauspartialisiert werden. Im Folgenden bezeichnen wir den Vektor der suffizienten Statistiken für die Nuisance-Parameter als W mit Realisation w.

Die bedingte Likelihood für den Antwortvektor $X_v = x_v$ nimmt bei dieser Nullhypothese die folgende Form an:

$$Pr\left(X_v = x_v\right) = \frac{\exp\left[\sum_i \left(\delta_{i2} \sum_{j \in I_i} x_{vj2}\right)\right]}{\sum \exp\left[\sum_i \left(\delta_{i2} \sum_{j \in I_i} x_{vj2}\right)\right]} \quad . \tag{6.26}$$

Ähnlich wie in Kapitel 5.3 können wir die Items, die zweimal oder gar nicht gelöst wurden, vernachlässigen. Mit n_{j_i} bezeichnen wir die Menge aller Items aus I_{j_i}, die genau einmal gelöst wurden. Die Zahl der nur im zweiten Messzeitpunkt gelösten Items aus

I_{j_i} bezeichnen wir als k_{j_i}. Da bei vorgegebenen Scoresummen $x_{v1\cdot}, \ldots, x_{vm\cdot}$ die Antwortmuster $(x_{v11}x_{v12})$,...,(x_{v11}, x_{v12}) der Items $1, \ldots, n$ unabhängig voneinander sind, gilt Unabhängigkeit auch zwischen Mengen I_{j_l}, I_{j_i} von Antwortmustern. Wir können daher I_{j_1} isoliert betrachten:

$$Pr\left(\sum_{l \in I_{j_1}} X_{vl2} = k_{j_1} \mid x_{v1\cdot}, \ldots, x_{vm\cdot}\right) = \frac{\binom{n_{j_1}}{k_{j_1}} \exp\left(k_{j_1}\delta_{j_1}\right)}{\sum \binom{n_{j_1}}{l} \exp(\delta_{j_1})} \quad . \tag{6.27}$$

Dies ist aber, wie in Kapitel 5.3 beschrieben, eine Binomialverteilung mit Parameter

$$p_{j_1} = \frac{\exp(\delta_{j_1})}{1 + \exp(\delta_{j_1})} \quad .$$

Daher ist die gemeinsame Verteilung der $\sum_{l \in I_{j_1}} X_{vl2}, \ldots, \sum_{l \in I_{j_s}} X_{vl2}$ das Produkt von s unabhängigen Binomialverteilungen mit Stichprobengrößen n_{j_i} und Erfolgswahrscheinlichkeiten p_{j_i}.

Insgesamt erhält man hier einen Test in einer $2 \times s$-Kontingenztafel mit vorgegebenen Randsummen $n_{j_1}, \ldots, n_{j_s}$. Als Testgröße kann man daher z.B. die Testgröße eines χ^2-Anpassungstests für die Nullhypothese

$$H_0 : Pr\left(\sum_{l \in I_{j_1}} X_{vl2}, \ldots, \sum_{l \in I_{j_s}} X_{vl2} \mid x_{v1\cdot}, \ldots, x_{vm\cdot}\right) = \prod_i \binom{n_{j_i}}{k_{j_i}} \left(p_{j_i}\right)^{k_{j_i}} \left(1 - p_{j_i}\right)^{n_{j_i} - k_{j_i}}$$

verwenden, also

$$TG = \frac{\sum_i \left(k_{j_i} - n_{j_i} p_{j_i}\right)^2}{n_{j_i} p_{j_i}} \quad . \tag{6.28}$$

Ein exakter Test kann in dieser Situation folgendermaßen konstruiert werden:

- Berechne für alle möglichen Antwortmuster die Testgröße TG.
- Sortiere die Testgröße.
- Berechne die Wahrscheinlichkeit, dass die Testgröße irgendeines Antwortmusters kleiner ist als die Testgröße der zu untersuchenden Person.

Bei großen Stichproben ist die Testgröße (unter der Nullhypothese) χ^2-verteilt mit s Freiheitsgraden.

Kapitel 7

Zusammenfassende Verfahren für Personenfittests

Bei den in Kapitel 5 vorgestellten Untersuchungsmöglichkeiten wurden nur Untersuchungsmöglichkeiten für Einzelpersonen vorgestellt. Beim Einsatz von Methoden der Veränderungsmessung liegen jedoch zumeist Ergebnisse für mehrere Personen vor, die meist zu Populationsaussagen zusammengefasst werden sollen. Außerdem birgt die mehrfache Anwendung von Personenfittests auf eine Substichprobe die Gefahr eines unkontrollierten Gesamtsignifikanzniveaus. Schließlich können die Ergebnisse der hier vorgestellten Einzeltests auch als Zwischenergebnisse gesehen werden: Oft wird man daran interessiert sein, zu erfahren, unter welchen zusätzlichen Bedingungen die Nullhypothese der Einzeltests besonders häufig abgelehnt wird. Methoden dieser Art werden in den nächsten Abschnitten vorgestellt.

Dabei werden wir Gebrauch von Verfahren aus der Metaanalyse machen. In der Metaanalyse wird (genau wie in unserem Fall) versucht, die Methoden vieler unterschiedlicher Einzeluntersuchungen zusammenzufassen. Anders als bei vielen Metaanalysen liegen in unserem Fall zumeist gut vergleichbare Daten vor: Falls mögliche Nebenbedingungen kontrolliert wurden, wurden die Daten aller Versuchspersonen unter gleichen oder zumindest ähnlichen Umständen gemessen. Dies ist bei vielen Metaanalysen nicht der Fall: Oftmals fließen in eine Metaanalyse Untersuchungen unterschiedlicher Reliabilität ein, deren Zusammenfassung zumindest fragwürdig erscheint. Zusammenfassend lässt sich daher sagen, dass unsere Ausgangslage der Idealsituation einer Metaanalyse nahe kommt (vgl. dazu z.B. [Fricke and Treinies, 1985]). Die Abkürzung TG steht im Folgenden für „Testgröße“. Darunter fällt jede der in dieser Arbeit vorgestellten Testgrößen.

7.1 Eine aus dem Binomialtest abgeleitete Gesamttestgröße für nichtrandomisierte Einzeltests

Im Falle nichtrandomiserter Tests lässt sich ein einfaches Gesamttestverfahren durchführen. Dazu nehmen wir an, dass wir bei N Versuchspersonen nichtrandomisierte Personenfittests durchgeführt haben. Ähnlich wie oben kann jeder dieser Personenfittests als eine Zufallsgröße $\Phi(TG)$ mit zwei Ausprägungen 0 und 1 aufgefasst werden, deren Verteilung (unter der Nullhypothese) bekannt ist:

$$Pr(\Phi = 1) = \alpha \ . \tag{7.1}$$

Unter der Nullhypothese ist zudem jeder Testausgang $\Phi(TG)$ einer Person v unabhängig identisch verteilt wie Φ. Die Zahl der Ablehnungen, zu denen die einzelnen Personenfittests führen, ist folglich binomialverteilt mit Parameter $p = Pr(\Phi = 1)$ und Stichprobengröße N. Daher ist eine Kontrolle des Gesamtsignifikanzniveaus möglich, indem wir die Nullhypothese

$$H_0 : Pr(\Phi(TG) = 1)) \leq \alpha \forall v$$

gegen die Alternative

$$H_A : Pr(\Phi(TG) = 1)) \geq \alpha \forall v$$

überprüfen. Dies kann mit dem Binomialtest geschehen, bei dem die (Gesamt-) Nullhypothese genau dann abgelehnt wird, wenn die Testgröße

$$N_k = \sum_v \Phi((TG))$$

größer als eine Zahl $k(\alpha^*)$ ist. Dabei ergibt sich $k(\alpha^*)$ als die kleinste ganze Zahl, die größer als das α^*-Quantil der $B(N; \alpha)$-Verteilung ist. Mit α^* wird hierbei das Signifikanzniveau des Gesamttests bezeichnet, das einen anderen Wert als das Signifikanzniveau α der Einzeltests annehmen darf. Bei großen Stichproben ist N_k approximativ normalverteilt mit Erwartungswert $N\alpha$ und Varianz $N\alpha(1 - \alpha)$.

Diese Methode kann nur dann angewendet werden, wenn sich das Ergebnis eines Einzeltests Φ als dichotome Größe darstellen lässt. Dies ist stets der Fall, wenn man nichtrandomisierte Einzeltests verwendet. Bei randomisierten Einzeltests besitzt die Zufallsgröße $\Phi(TG)$ hingegen 4 verschiedene Ausprägungen. Um das Binomialtestverfahren anwenden zu können, muss im Fall randomisierter Einzeltests daher eine Größe $\Phi^*(TG)$ verwendet werden, die mit $\Phi(TG)$ folgendermaßen verbunden ist:

$$\Phi^*(TG) = \begin{cases} \Phi(TG) & \text{falls} \quad TG \in K_0 \cup K_A \\ Z_{v1} & \text{falls} \quad TG = C_1 \\ Z_{v2} & \text{falls} \quad TG = C_2 \end{cases} . \tag{7.2}$$

Dabei ist Z_{vi} eine dichotome Zufallsgröße, die den Wert 1 mit Wahrscheinlichkeit γ_i annimmt, sowie den Wert 0 mit Wahrscheinlichkeit $1-\gamma_i$. C_1, C_2 sind die Grenzen zwischen dem Annahmebereich K_0 und dem Ablehnbereich K_A.

7.2 Klassische Testgrößen der Metaanalyse

Die klassischen Methoden der Metaanalyse für die inferenzstatistische Beurteilung von P-Werten sind als „Adding of logs“ (vgl. [Klauer, 1991b], [Fricke and Treinies, 1985], Kap. 5.2.) bzw. „Signifikanzprüfung für einen mittleren P-Wert“ (vgl. [Fricke and Treinies, 1985], Kap 5.2.) bekannt.

Adding of logs

Dieses Verfahren beruht auf der Tatsache, dass unter recht allgemeinen Voraussetzungen (Unabhängigkeit der einzelnen Testgrößen) die globale Testgröße

$$G^2 = \sum_v -2log[\kappa(TG)] \tag{7.3}$$

eine χ^2-Verteilung mit $2N$ Freiheitsgraden besitzt (vgl. dazu auch z.B. [Lienert, 1973], S. 509ff). Dabei ist $\kappa(TG)$ der P-Wert des Personenfittests für die Person v. Die vorgestellte Testgröße erlaubt einen Gesamttest für alle Personen einer Stichprobe.

Signifikanzprüfung für einen mittleren P-Wert

Bei ausreichender Stichprobengröße n (laut [Fricke and Treinies, 1985]: $n > 4$) und Unabhängigkeit zwischen den Versuchspersonen ist

$$Z = (0.5 - \bar{\kappa}(TG))\sqrt{12n} \tag{7.4}$$

$N(0,1)$-verteilt (vgl. [Fricke and Treinies, 1985]). Dieser Zusamenhang lässt sich somit leicht für die Bestimmung eines approximativen Gesamtsignifikanzniveaus verwenden, wobei eine Nullhypothese über den mittleren Wert von $\kappa(TG)$ benutzt wird. Wie oben bezeichnet $\kappa(TG)$ auch hier den P-Wert des Einzeltests bei Person v.

7.3 Diskussion der Testmethoden für eine Zusammenfassung der Personenfittests

An dieser Stelle soll die Verwendung der vorgestellten Gesamttestverfahren diskutiert werden. Kritischer Punkt beim Einsatz aller vorgestellten Methoden ist die Frage der Unabhängigkeit der P-Werte $\kappa(TG)$ (bzw.: der Zufallsgrößen $\Phi(TG)$) einzelner Personen. Im Rasch-Modell ist diese Art von Unabhängigkeit aber durch die Annahme der „lokalen stochastischen Unabhängigkeit“ gewährleistet.

Die Entscheidung zwischen einer auf $\Phi(TG)$ beruhenden vs. einer auf $\kappa(TG)$ aufbauenden Gesamttestgröße ist weniger eindeutig zu fällen. Für die auf $\kappa(TG)$ aufbauenden Testgrößen spricht, dass die in der Stichprobe anfallende Information besser verwertet wird. $\kappa(TG)$ enthält Information über die Stärke der Abweichung zur Nullhypothese, die bei der Verwendung der dichotomen Größe $\Phi(TG)$ verloren geht.

Für eine auf $\Phi(TG)$ aufbauende Testgröße spricht hingegen, dass wesentlich klarer definiert werden kann, wann in einer Stichprobe von schlechtem Personenfit gesprochen werden kann: In die Gesamttestgröße N_k gehen nur solche Personen ein, bei denen die Einzelnullhypothese abgelehnt werden kann, d.h. die schlecht angepassten Personen. In eine Testgröße, die $\kappa(TG)$ verwendet, gehen hingegen alle Personen einer Stichprobe ein, deren P-Wert kleiner als 1 ist. I.d.R. sind dies auch viele Personen, bei denen es nicht zur Ablehnung der Einzelnullhypothese kommt. Dies wiederum kann zu der paradoxen Situation führen, dass zwar kein Einzeltest zur Ablehnung der Einzelnullhypothese führt, jedoch der Gesamttest zur Ablehnung der Gesamtnullhypothese.

Der Einsatzbereich der auf $\kappa(TG)$ aufbauenden Gesamttestgröße ist eher bei Fragestellungen zu sehen, bei denen die mittlere Personenfitabweichung einer Stichprobe beobachtet werden soll. Die auf $\Phi(TG)$ aufbauenden Tests scheinen dagegen zielgerichteter für die Absicherung eines Gesamtsignifikanzniveaus bei reinen Personenfituntersuchungen zu sein.

7.4 Deskriptive Methoden der Zusammenfassung einzelner Ergebnisse: Effektgrößen

In diesem Abschnitt wird die Zusammenfassung der Personenfittests mittels Effektgrößen behandelt. Es wird die Effektgröße für den Binomialtest als Maßstab für die Größe des aufgetretenen Effektes eingeführt. Diese Methode kann nur jeweils für den Effekt eines einzelnen Traits bei 2 Zeitpunkten verwendet werden.

Als Effektgröße ist im hier vorliegenden Fall des Vergleichs eines geschätzten Anteilswertes $\hat{p}$ mit einem (durch die Nullhypothese) vorgegebenen Wert π_0 die folgende Effektgröße anzuwenden:

$$\Delta = \arcsin(\sqrt{\hat{p}}) - \arcsin(\sqrt{\pi})) \tag{7.5}$$

(vgl. auch [Cohen, 1969], Kap. 6). $\hat{p}$ ist in diesem Fall folgendermaßen definiert: Bei Trait I_l wurden m Items in beiden Zeitpunkten unterschiedlich beantwortet. Von diesen wurden k Items im zweiten, nicht aber im ersten Zeitpunkt beantwortet. Wir berechnen dann

$$\hat{p} = \frac{k}{n_1} \quad , \tag{7.6}$$

sowie

$$\pi_0 = \frac{\exp\left(\delta_{g2}^0\right)}{1 + exp(\delta_{g2}^0)} \quad . \tag{7.7}$$

Hierbei ist m_i die Zahl der nur in einem Zeitpunkt richtig beantworteten Items bei Person i. Bei m_i in den Binomialtest eingehenden Items ist $\Delta\sqrt{m_i}$ standardnormalverteilt (vgl. [Cohen, 1969], Kap. 6.5). Dies kann man verwenden, um einen Test auf Signifikanz des Gesamteffekts zu konstruieren: Die Testgröße

$$TG = \sum_i m_i \Delta_i^2 \tag{7.8}$$

ist χ^2-verteilt mit N Freiheitsgraden, falls man annimmt, dass die Effektgrößen der einzelnen Personen unabhängig voneinander sind. Hieraus kann ein Testverfahren für die Gesamtstichprobe konstruiert werden. $i \in \{1, \ldots, N\}$ ist dabei der Index für Versuchspersonen, N die Gesamtzahl der Versuchspersonen, für die ein Personenfittest durchgeführt wurde. Nullhypothese ist dabei:

$$E(\Delta_1) = \ldots = E(\Delta_N) = 0 \quad ,$$

was gegen die Alternative:

$$E\Delta_i \neq 0 \text{ für mindestens eine Versuchsperson i}$$

getestet wird. Bei Gültigkeit der Alternativhypothese ist TG nichtzentral χ^2-verteilt mit Zentralitätsparameter $\sum (E\Delta_i)^2$, da laut [Witting, 1978] die Summe der Quadrate mehrerer unabhängiger $N(\mu, \sigma)$-verteilter Zufallsgrößen nichtzentral χ^2-verteilt ist.

Falls die Nullhypothese nicht zutrifft, nimmt TG somit große Werte an. Den Ablehnbereich dieses Gesamttests erhält man daher aus dem $1 - \alpha$-Quantil der χ^2-Verteilung mit N Freiheitsgraden. Neben diesem Gebrauch als Signifikanztest für die gesamte Stichprobe kann man

$$\overline{\Delta} = \sum (\Delta_i \sqrt{m_i}) / N$$

aber auch als deskriptives Maß für den mittleren Veränderungseffekt verwenden.

$\Delta\sqrt{m_i}$ kann darüber hinausgehend auch zur Planung der Stichprobengröße verwendet werden. Wenn in Wirklichkeit ein Wert π_A statt π_0 der wahre zu schätzende Anteilswert ist, dann ist

$$\Delta\sqrt{m_i} = \left(\arcsin(\sqrt{\hat{p}}) - \arcsin(\sqrt{\pi_A})\right)\sqrt{m_i}$$

eine approximativ $N(\mu, 1)$-verteilte Zufallsgröße, wobei

$$\mu = \arcsin(\sqrt{\pi_A})) - \arcsin(\sqrt{\pi_0})) \tag{7.9}$$

ist. Mit Hilfe der $N(\mu, 1)$-Verteilung kann die Stichprobengröße geplant werden, indem man die kleinste Stichprobengröße wählt, bei der ein Mindestwert für $|\mu|$ bei vorgegebenem Mindestsicherheitsgrad γ und vorgegebenem Gesamtsignifikanzniveau α gerade noch erkannt werden kann. Dies sind Standardmethoden der Stichprobenplanung, vgl. z.B. [Cohen, 1969]. Eine solche Vorgehensweise bietet sich z.B. beim Einsatz des LLTM zur Evaluation von Psychotherapien an (vgl. Kapitel 2.1 und 12.5).

Abschließend bleibt anzumerken, dass die hier vorgeschlagenen Effektgrößen auch für den in Kapitel 6.3 als Personenfittest vorgeschlagenen exakten Test von Fisher verwendet werden können.

Kapitel 8

Tests bei bekannten Itemparametern

In diesem Kapitel beschreiben wir Tests, bei denen die Kenntnis der Schwierigkeitsparameter der Items angenommen wird. Dies sind einerseits Tests für den Fall, dass unterschiedliche Items an den beiden Messzeitpunkten verwendet werden, andererseits Tests zur Aufdeckung des Phänomens „Erinnerung an den vorhergehenden Messzeitpunkt“. Vorausgesetzt wird dabei immer ein LLTM mit dichotomen Items und zwei Messzeitpunkten.

8.1 Gleichmäßig beste Tests bei unvollständiger Itemwiederholung

Einleitung

Die bisher vorgestellten gleichmäßig besten (unverfälschten) Tests gehen davon aus, dass eine Menge von Items geschlossen an zwei Zeitpunkten dargeboten wird. Dies ist jedoch oftmals eine unrealistische Annahme, die zudem für systematische Fehler durch Erinnerungseffekte im zweiten Zeitpunkt verantwortlich sein kann (vgl. [Embretson, 1991]). Daher ist es für die praktische Anwendung von Tests auf Personenfit wichtig, dass man auch kompliziertere Untersuchungsdesigns ermöglicht. Dies kann z.B. dadurch geschehen, dass man Ankeritems verwendet, die zu beiden Zeitpunkten dargeboten werden, während die anderen Items nur an jeweils einem Zeitpunkt abgefragt werden (vgl. z.B. [Fischer, 1995b]). Eine extreme Möglichkeit dieses Vorgehens bestünde darin, zu den beiden Zeitpunkten nur verschiedene Items darzureichen, wobei (wenn wir für den Augenblick von nur zwei Untersuchungsgruppen ausgehen) die Versuchsgruppe 2 im ersten Zeitpunkt die Items verwendet, die Versuchsgruppe 1 im zweiten Zeitpunkt benützt (sowie vice versa).

Dies lässt sich in der folgenden Tabelle ausdrücken:

	Gruppe 1	Gruppe 2
Zeitpunkt 1	Items 1,...,m Items 2m+1,...,3m	Items 1,...,m Items m+1,...,2m
Zeitpunkt 2	Items m+1,...,2m Items 3m+1,...,4m	Items 2m+1,...,3m Items 3m+1,...,4m

Zu diesem Design: vgl. z.B. [Fischer, 1995c] oder [Embretson, 1991].

Das im vorhergehenden Kapitel vorgeschlagene Design kann jetzt nicht mehr verwendet werden, da hier keine Partialsummen über die Zahl der Lösungen eines Items bei einer bestimmten Person mehr berechnet werden können. Eine Lösung dieses Problems ergibt sich, wenn man die Itemschwierigkeiten im ersten Zeitpunkt als gegeben betrachtet. Dieses Vorgehen wird in der gesamten Literatur zu Personenfittests vorgeschlagen (u.a. [Klauer, 1995], [Molenaar and Hoijtink, 1990]). [Fischer, 1995c] geht bei der Messung von Modifiability auf gleiche Weise vor.

Herleitung der Testgröße und ihrer Verteilung

Im Folgenden leiten wir eine Testgröße für den Fall unvollständiger Itemwiederholung her. Die hierbei vorgestellten Ergebnisse folgen auch aus dem in [Fischer, 1995c] vorgestellten Modell zur Messung von Modifiability.

Falls wir die Itemschwierigkeiten als gegeben voraussetzen können, erhalten wir für das Testergebnis einer Person v eine andere Likelihood:

$$\begin{aligned} P(X_v = x_v|\beta_i) &= \\ &= \frac{exp[\sum_{i,t} x_{vit}(\theta_v - \beta_i + \delta_{gt})]}{\prod_{i,t}[1 + exp(\theta_v - \beta_i + \delta_{gt})]} = \\ &= \prod_{i,t} C^{-1}(\theta_v, \beta_i, \delta_{gt}) \times exp\left[\sum_i -x_{vi.}\beta_i\right] \times exp\left[x_{v..}\theta_v - x_{v.2}\delta_{g2}\right] \quad . \end{aligned} \tag{8.1}$$

Dabei wird mit $x_{vi.}$ die Summe der Ergebnisse der Person v bei Item i über alle Zeitpunkte bezeichnet, an denen Item i der Person v vorgelegt wurde. Dies ist die Likelihood einer zweiparametrigen Exponentialfamilie mit natürlichen Parametern θ_v und δ_{g2} sowie suffizienten Statistiken $x_{v..}$, sowie $x_{v.2}$. Für diese zweiparametrige Exponentialfamilie kann leicht ein gleichmäßig bester (unverfälschter) Test konstruiert werden, der nach denselben Prinzipien funktioniert wie der zuerst vorgestellte Test.

Als bedingte Verteilung unter der Bedingung $X_{v\cdot\cdot} = x_{v\cdot\cdot}$ erhält man:

$$Pr(X_v = x_v | x_{v\cdot\cdot}, \beta_i) = exp[-\sum_i x_{vi\cdot}\beta_i] \times \frac{exp[x_{v\cdot 2}\delta_{g2}]}{\gamma_{x_{v\cdot\cdot}}} \quad . \tag{8.2}$$

Dabei ist $\gamma_{x_{v\cdot\cdot}}$ die zu den Parameterschwierigkeiten $\beta_1, ..., \beta_m$ gehörende elementare symmetrische Funktion für Summenscore $x_{v\cdot\cdot}$.

Im Folgenden nehmen wir o.B.d.A. an, dass zum ersten Messzeitpunkt die ersten k Items mit Schwierigkeiten $\beta_1, ..., \beta_k$ sowie zum zweiten Zeitpunkt die l letzten Items mit Schwierigkeiten $\beta_l, ..., \beta_m$ dargeboten wurden, wobei $k \leq m$ und $l \leq m$ gilt. Dann erhalten wir mit $r_{min} = max(0, x_{v\cdot\cdot} - m)$, sowie mit $r_{max} = min(m, x_{v\cdot\cdot})$:

$$\begin{aligned}
&Pr(X_{v\cdot 2} = x_{v\cdot 2} \mid \beta_1, ..., \beta_m, x_{v\cdot\cdot}) = \\
&= \frac{\sum_{x_v | X_{v\cdot 2} = x_{v\cdot 2}} P(X_v = x_v)}{P(X_{v\cdot\cdot} = x_{v\cdot\cdot})} = \\
&= \frac{\sum_{x_v | X_{v\cdot 1} = x_{v\cdot 1}} [exp(\sum_i x_{vi1}\beta_i)] \sum_{x_v | X_{v\cdot 2} = x_{v\cdot 2}} [exp(\sum_i x_{vi2}\beta_i)] exp(-x_{v\cdot 2}\delta_{g2})}{\sum_{l=r_{min}}^{r_{max}} \gamma^{t=1}_{x_{v\cdot\cdot}-l} \gamma^{t=2}_{l} exp(-l\delta_{g2})} = \\
&= \frac{\gamma^{t=1}_{x_{v\cdot 1}} \times \gamma^{t=2}_{x_{v\cdot 2}} \times exp(-x_{v\cdot 2}\delta_{g2})}{\sum_{l=r_{min}}^{r_{max}} \gamma^{t=1}_{x_{v\cdot\cdot}-l} \gamma^{t=2}_{l} exp(-l\delta_{g2})} \quad .
\end{aligned} \tag{8.3}$$

Dies ist die Wahrscheinlichkeit des Auftretens für einen beliebigen Wert $x_{v\cdot 2}$ der für δ_{g2} suffizienten Statistik $X_{v\cdot 2}$. Da diese Wahrscheinlichkeit relativ leicht zu berechnen ist, ist die Entwicklung eines gleichmäßig besten unverfälschten Tests ab diesem Punkt „straightforward“ und entspricht der Vorgehensweise im Kapitel 5.2.

Das gerade vorgestellte Verfahren kann auch angewendet werden, wenn in beiden Messzeitpunkten die gleichen Items verwendet werden. Daher wird diese Testprozedur jetzt mit dem Vorgehen in Kapitel 5.2 verglichen. Vorteile des gerade besprochenen Verfahrens sind:

- Das Verfahren funktioniert bei komplizierten Itemanordnungen, insbesondere wenn an beiden Zeitpunkten unterschiedliche Items dargeboten werden. Dies ist vor allem im Hinblick auf das Phänomen „Erinnern an den ersten Zeitpunkt“ sinnvoll. Es ist auch möglich, an beiden Messzeitpunkten unterschiedliche Itemzahlen zu verwenden.
- Es gehen – anders als bei den Tests aus 5.2 – stets alle Items in die Testgrößenverteilung ein.

Das in Kapitel 5.2 vorgestellte Verfahren besitzt hingegen folgende Vorteile:

- Das Verfahren aus 5.2 vermeidet es, geschätzte Parameter als bekannt annehmen zu müssen.
- Das Verfahren aus 5.2 ist auch im LLRA anwendbar, was bei dem in diesem Kapitel vorgestellten Verfahren nicht der Fall ist.

Eine Effektgröße für Tests mit bekannten Itemschwierigkeiten

Analog zu Kapitel 7.4 geben wir auch für die Tests bei bekannten Itemparametern standardisierte Größen an, mit denen Vergleiche zwischen verschiedenen Personen bzw. Subpopulationen möglich sind. Dazu benutzen wir die Tatsache, dass die Verteilung (8.3) zur Familie der Generalized Power Series Distributions (= GPSD) gehört (vgl. z.B. [Fischer, 1995c]). Im Folgenden fassen wir einige wichtige Eigenschaften der GPSD zusammen (nach [Patel et al., 1976]):

- Die GPSD sind definiert auf einer nichtleeren Trägermenge $T \in \mathbf{N}_0$ (Anmerkung: $\mathbf{N}$ steht für die Menge der natürlichen Zahlen). Mit $a(x)$ bezeichnen wir positive auf T definierte Funktionen $a : T \mapsto \mathbf{N}$. Den reellwertigen Parameter der GPSD bezeichnen wir mit η. Dann besitzen die GPSD folgende Wahrscheinlichkeitsfunktion für $x \in T$:
$$P(X = x) = \frac{a(x)\,\eta^x}{\sum_{x \in T} a(x)\,\eta^x} \ . \tag{8.4}$$
- Erwartungswert: $EX = \eta \frac{d}{d\eta}[\log(\sum_{x \in T} a(x)\,\eta^x)]$
- Varianz: $Var(X) = EX + \eta^2 \frac{d^2}{d\eta^2}[\log(\sum_{x \in T} a(x)\,\eta^x)]$
- EX ist eine nichtnegative monoton steigende Funktion von η.
- Falls eine Verteilung zu den GPSD gehört, so ist sie durch die Angabe der ersten beiden nichtzentralen Momente eindeutig bestimmt.

Aus der Nichtnegativität von EX sowie der Identifizierbarkeit einer GPSD-Verteilung durch die ersten beiden Momente folgt, dass eine GPSD-Verteilung auch durch Angabe von Varianz und Erwartungswert eindeutig identifizierbar ist. Letzteres folgt aus der Bijektivität der Funktion

$$f\left(EX, EX^2\right) = (EX, Var(X)) \quad , \tag{8.5}$$
$$f : ([0;\infty] \times [0;\infty]) \mapsto ([0;\infty] \times [0;\infty]) \quad .$$

Für jeden Wert der Testgröße $X_{v\cdot 2} = x_{v\cdot 2}$ kann ein standardisierter Wert

$$\Delta = \frac{x_{v\cdot 2} - E_{\delta^0_{g2}} X_{v\cdot 2}}{\sqrt{Var_{\delta^0_{g2}}(X_{v\cdot 2})}} \tag{8.6}$$

berechnet werden, der Vergleiche bezüglich der Abweichung von der Nullhypothese in Abhängigkeit vom Träger T möglich macht. Δ ist der standardisierte Abstand der Realisation $x_{v\cdot 2}$ der Testgröße von dem unter der Nullhypothese zu erwartenden Wert $E_{\delta^0_{g2}} X_{v\cdot 2}$, und ermöglicht einen Vergleich auch zwischen GPSD-Verteilungen mit unterschiedlichen Trägermengen resp. unterschiedlichen Gewichtsfunktionen $a(x)$.

Eine zweite Möglichkeit, mit der die Testgrößen $X_{v\cdot 2}$ untereinander vergleichbar gemacht werden können, besteht in der Anwendung einer varianzstabilisierenden Transformation $L(X)$, die dafür sorgt, dass die Varianz der Verteilung bei wachsendem Erwartungswert stabil bleibt. [Falk et al., 1995] Kap. 1.7. folgend, erhält man für die Transformation $L(X)$ mittels einer Taylor-Entwicklung:

$$\begin{aligned} Var(L(X)) &= c \approx \text{const} \\ &\approx \left(L^{'}\left(E_\eta X\right)\right)^2 V(E_\eta X) \ . \end{aligned} \tag{8.7}$$

Dies gilt für beliebige Zufallsgrößen X mit Parameter η und Varianzfunktion $V(E_\eta X)$. Im Folgenden setzen wir

$$b = \eta^2 \frac{d^2}{d\eta^2} [\log(\sum_{x \in T} a\left(x\right) \eta^x)$$

sowie $Var(X) = E_\eta X + b$ in die obige Gleichung ein. Dadurch erhalten wir

$$\left(L^{'}\left(E_\eta X\right)\right)^2 V(E_\eta X) = \left(L^{'}\left(E_\eta X\right)\right)^2 (E_\eta X + b) = c,$$

was – falls $V(E_\eta X) > 0$ und $Var(L(X)) > 0$ gelten – zu

$$L^{'}\left(E_\eta X\right) = \frac{c}{\sqrt{E_\eta X + b}}$$

und somit zu

$$L\left(E_\eta X\right) = c \int \frac{1}{\sqrt{E_\eta X + b}} = 2c\sqrt{E_\eta X + b}$$

führt. Daher ist durch

$$L\left(X\right) = 2\sqrt{E_\eta X + b} \tag{8.8}$$

eine varianzstabilisierende Transformation gegeben, die beim Vergleich von Effekten verwendet werden kann.

Um zu überprüfen, ob ein Gesamteffekt für die Stichprobe vorliegt empfiehlt sich unserer Auffassung nach ein Likelihood-Ratio-Test. Dazu sollte die Likelihood (8.3) verwendet werden. (8.3) ist die Wahrscheinlichkeit einer Person, einen bestimmten Summenscore im zweiten Messzeitpunkt zu erzielen, falls der Gesamtsummenscore vorgegeben ist.

Die Häufigkeit, mit der Personen aus Untersuchungsgruppe g mit Gesamtsummenscore $x_{v\cdot\cdot}$ einen Teilsummenscore $x_{v\cdot 2}$ erzielen, kann in einer $(m^2+1) \times m \times G$-Kontingenztafel dargestellt werden. Hierbei steht G für die Gesamtzahl der Untersuchungsgruppen, sowie m

für die Zahl der Items pro Untersuchungszeitpunkt. In dieser Kontingenztafel überprüft man nun, ob die aufgetretenen Häufigkeiten der einzelnen Zellen durch (8.3) erklärbar sind. Die Testgröße des Likelihood-Ratio-Tests ist dann – unter der Nullhypothese, dass die durch das Rasch-Modell geschätzten Veränderungsparameter zutreffen – χ^2-verteilt mit $(m^2G\,(m-1)) - G - m^2 - m$ Freiheitsgraden. Diese Zahl von Freiheitsgraden gilt, wenn man die Schwierigkeitsparameter der Items als gegeben ansieht. Falls die Schwierigkeitsparameter geschätzt wurden, verringert sich die Zahl der Freiheitsgrade um die Zahl der geschätzten Itemparameter.

Beispiel

Um die Funktionsweise der in diesem Abschnitt vorgestellten Tests zu erläutern, rechnen wir ein Anwendungsbeispiel. Der vorgegebene Datensatz enthält 2 Gruppen zu je 15 Versuchspersonen, denen an 2 Zeitpunkten je 11 Items dargeboten werden. Die tatsächlichen Antworten werden folgendermaßen simuliert:

- In beiden Gruppen werden die Werte $-2, -1.6, -1.2, \ldots, -0.4, 0, 0.4, \ldots, 2$ als Itemschwierigkeiten verwendet, als Fähigkeitsparameter die Werte $-2, -1, 0, 1, 2$.
- In Gruppe 1 werden zu beiden Zeitpunkten Itemantworten gemäß einem LLTM mit den o.g. Itemschwierigkeiten resp. Fähigkeiten sowie dem Veränderungsparameter 0.5 simuliert.
- In Gruppe 2 werden die Antworten im 1. Messzeitpunkt gemäß dem Rasch-Modell mit den o.g. Itemschwierigkeiten resp. Fähigkeiten modelliert. Im zweiten Messzeitpunkt wird ein Erinnerungsmodell angenommen: Mit Wahrscheinlichkeit 0.7 wird die gleiche Antwort wie im ersten Messzeitpunkt gegeben, andernfalls gemäß dem LLTM eine Antwort simuliert. Die Veränderung wird mit dem Wert 0 simuliert.
- In beiden Gruppen wird bei jeweils 3 Personen das Antwortmuster im zweiten Messzeitpunkt mit einer Veränderung vom Wert -3 simuliert. Die hierbei ausgewählten Personen besitzen die Fähigkeitswerte -2 resp. 0 resp. 2.

Anschließend werden Personenfittests durchgeführt, einmal mit der Annahme bekannter Itemschwierigkeiten, einmal ohne diese Annahme. Es werden P-Werte für die einzelnen Personen berechnet. Die P-Werte werden dabei als die Auftrittswahrscheinlichkeit α des kleinsten Ablehnbereichs eines randomisierten Tests definiert, bei dem der für eine Person gemessene Testgrößenwert ganz im Ablehnbereich liegt.

In Gruppe 1 zeigt sich Folgendes: Die Testvariante ohne Kenntnis der Itemschwierigkeiten kann bei allen 15 Versuchspersonen durchgeführt werden. Bei 5 Versuchspersonen unterscheiden sich die Antwortmuster an den beiden Zeitpunkten nur in ein oder zwei Items. In diesen Fällen erscheint die Durchführung eines Personenfittests wegen zu geringer Stichprobengröße wenig sinnvoll. Bei den übrigen 7 Versuchspersonen mit simulierter Veränderung 0.5 ergeben sich P-Werte zwischen 0.142 und 1. Für die Personen mit

Veränderung -3 und Fähigkeit -2 resp. 0 resp. 2 ergeben sich die P-Werte > 0.999 resp. 0.377 resp. 0.020. Bei einem vorgegebenen Signifikanzniveau von $\alpha = 0.05$ wäre also nur eine von drei aberranten Versuchspersonen erkannt worden.

Gruppe	P-Wert	Anzahl
1	0.020	1
	0.142	3
	0.350	3
	0.377	2
	0.494	1
	0.612	1
	>0.999	4
2	0.500	1
	>0.999	7

Tabelle 8.1: Häufigkeit einzelner P-Werte bei der Testvariante ohne Itemparameter. Gliederung nach Untersuchungsgruppe.

Bei der Testvariante mit Kenntnis der Itemparameter kann der Test in Gruppe 1 ebenfalls bei allen Versuchspersonen durchgeführt werden. Bei den 12 Versuchspersonen mit simulierter Veränderung 0.5 ergeben sich P-Werte zwischen 0.172 und 0.999. Von den 3 aberranten Versuchspersonen kann bei Signifikanzniveau $\alpha = 0.05$ nur die Versuchsperson mit Fähigkeit 2 entdeckt werden (mit P-Wert 0.038). Bei den beiden anderen Versuchspersonen ergeben sich P-Werte von 0.670 resp. > 0.999.

Gruppe	P-Wert	Anzahl
1	0.038	1
	0.172	1
	0.245	1
	0.281	1
	0.303	1
	0.539	1
	0.606	2
	0.670	1
	>0.999	6
2	0.606	1
	>0.999	14

Tabelle 8.2: Häufigkeit einzelner P-Werte bei der Testvariante mit bekannten Itemparametern. Gliederung nach Untersuchungsgruppe.

In Gruppe 2 kann das Verfahren ohne Itemparameter nur bei 8 von 15 Versuchspersonen durchgeführt werden, da sich bei den anderen Personen an beiden Zeitpunkten die gleichen Antwortmuster ergeben. Das Verfahren mit bekannten Itemparametern kann bei allen Versuchspersonen durchgeführt werden. Bis auf eine Testperson erhalten wir in

Gruppe 2 bei beiden Testverfahren immer einen P-Wert > 0.999. Somit können die aberranten Personen durch beide Testverfahren nicht entdeckt werden.

Abschließend bleibt zu sagen, dass sich die beiden Testvarianten zumindest in dem verwendeten Datensatz ähnlich verhalten. Bei beiden Testverfahren zeigt sich eine starke Abhängigkeit vom Ausgangswert. Dieser Befund wird sich ähnlich in den nächsten Kapiteln zeigen. Weiterhin können durch das Auftreten des Störfaktors „Erinnerung“ vorhandene aberrante Antwortmuster nicht entdeckt werden. Zumindest in diesem Beispiel wird durch „Erinnerung“ die praktische Anwendbarkeit der in dieser Arbeit vorgestellten Tests deutlich verringert.

Am Ende dieses Abschnitts sei angemerkt: Auch dieses Modell kann auf den Fall polytomer Items verallgemeinert werden. Als Beispiel dazu setzen wir als polytomes Rasch-Modell ein LPCM voraus. Weiterhin setzen wir voraus, dass es genau einen Veränderungsparameter δ gibt, so dass folgendes Modell gilt:

$$\begin{aligned} &P(\mathbf{X}=\mathbf{x}|\theta_v,\tau_{111},\ldots,\tau_{2mJ}) = \\ &= \frac{\exp\left(x_{v\cdot 1}\theta_v - \sum_{i=1}^{m}\sum_{h=1}^{x_{vi1}}\tau_{1ih} + x_{v\cdot 2}\theta_v - \sum_{i=1}^{m}\sum_{h=1}^{x_{vi2}}\tau_{1ih} + \sum_{i=1}^{m}x_{2vi}\delta_{2g}\right)}{\prod_{t=1}^{2}\prod_{i=1}^{m}\sum_{j=0}^{J}\exp\left(j\theta_v - \sum_{h=1}^{j}\tau_{1ih}\right)} \quad . \end{aligned} \tag{8.9}$$

Man erkennt leicht, dass auch in diesem Fall ein Test mit $TG(x) = \sum_{i=1}^{m} x_{2vi}$ als Testgröße konstruiert werden kann. Auch in diesem Fall liegt eine 2-parametrige Exponentialfamilie vor, so dass wir auf ein ähnliches Ergebnis wie in Gleichung (8.3) kommen.

8.2 Test auf Erinnern

Der in diesem Abschnitt vorgestellte Test modelliert das Phänomen „Erinnern“ als Abweichung von der lokalen stochastischen Unabhängigkeit mit Hilfe eines loglinearen Modells. Ein Spezialfall dieses loglinearen Modells (falls keine „Erinnerung“ vorliegt) ist das in den bisherigen Kapiteln verwendete LLTM. Die Antwortwahrscheinlichkeit für den 2. Messzeitpunkt modelliert sich mittels

$$Pr(X_{vi2} = x_{vi2} \mid x_{vi1}) = \frac{\exp\left[x_{vi2}\left(\theta_v - \beta_i + \delta_2\right) + z_{vi}\lambda\right]}{1 + \exp\left[x_{vi2}\left(\theta_v - \beta_i + \delta_2\right) + z_{vi}\lambda\right]} \quad , \tag{8.10}$$

wobei gilt:

$$z_{vi} = -x_{vi2}x_{vi1} + \left(1 - x_{vi1}\right)\left(1 - x_{vi2}\right) \quad . \tag{8.11}$$

Durch z_{vi} wird eine Verletzung der lokalen stochastischen Unabhängigkeit zwischen den beiden Zeitpunkten modelliert: Falls $\lambda > 0$ gilt, erhöht sich die Wahrscheinlichkeit gleicher Antworten an beiden Messzeitpunkten, bei $\lambda < 0$ wird diese Wahrscheinlichkeit niedriger. Somit ist dies eigentlich kein Test auf Erinnern, sondern auf den Zusammenhang zwischen 2 Messzeitpunkten. Falls $\lambda = 0$ gilt, wird dieses Modell zu einem LLTM für die Veränderungsmessung. Modelle dieser Art werden in der Klasse der loglinearen Rasch-Modelle zusammengefasst (zu den loglinearen Rasch-Modellen: vgl. [Kelderman, 1984], [Klauer, 1995], [Ponocny, 2000]).

Unter den Nebenbedingungen

- Antwortmuster im ersten Messzeitpunkt bekannt,
- Itemparameter β_i bekannt,
- Summenscore im zweiten Messzeitpunkt ist suffiziente Statistik für δ_2

liegt für den Antwortvektor $X_{v2} = (X_{v12}, \ldots, X_{vm2})$ mit Realisation $x_{v2} = (x_{v12}, \ldots, x_{vm2})$ eine einparametrige Exponentialfamilie in λ vor:

$$\begin{aligned} &Pr\left(X_{v2} = x_{v2} \mid \beta_1, \ldots, \beta_m, x_{v11}, \ldots x_{vm1}, x_{v\cdot 2}\right) = \\ &= \frac{\exp\left(\sum\left(x_{vi2}\beta_i\right) + \lambda z_{v\cdot}\right)}{\sum_{X_{v2}\in\mathcal{I}} \exp\left(\sum\left(x_{vi2}\beta_i\right) + \lambda z_{v\cdot}\right)} \quad . \end{aligned} \tag{8.12}$$

Hierbei ist $\mathcal{I}$ die Menge aller Antwortmuster, für die $X_{v\cdot 2} = x_{v\cdot 2}$ gilt, wenn das Antwortmuster im ersten Messzeitpunkt die Werte $x_{v11}, \ldots x_{vm1}$ annimmt.

Folglich existiert zu der Nullhypothese

$$\lambda \geq \lambda_0$$

bei der Alternativhypothese

$$\lambda < \lambda_0$$

ein gleichmäßig bester Test mit Testgröße $Z_{v\cdot} = \sum Z_{vi}$, da hier eine einparametrige Exponentialfamilie in $Z_{v\cdot}$ vorliegt (zur einparametrigen Exponentialfamilie: vgl. auch [Witting, 1978], S. 84ff.). Die Testgröße $Z_{v\cdot}$ besitzt die Verteilung

$$Pr\left(Z_{v\cdot} = Z_{v\cdot} \mid X_{v2} \in \mathcal{I}\right) = \frac{\sum \exp\left(\sum_i \left(x_{vi2}\beta_i\right) + \lambda z_{v\cdot}\right)}{\sum_{X_{v2}\in\mathcal{I}} \exp\left(\sum\left(x_{vi2}\beta_i\right) + \lambda z_{v\cdot}\right)} \quad . \tag{8.13}$$

Dabei wird im Zähler dieses Bruchs über alle Antwortmuster aus $\mathcal{I}$ summiert, bei denen $Z_{v\cdot} = z_{v\cdot}$ zutrifft. Nun wird ein Algorithmus zur Berechnung der Verteilung von $Z_{v\cdot}$ vorgestellt. Vorauszusetzen ist, dass

- man das Antwortmuster x_{v1} im ersten Messzeitpunkt und somit auch den Summenscore $x_{v\cdot 1}$ kennt,
- der Summenscore $x_{v\cdot 2}$ des zweiten Messzeitpunkts vorgegeben ist, sowie dass
- die Itemschwierigkeiten β_i bekannt sind.

Weiterhin verwenden wir die Bezeichnung

$$\kappa(z_{v\cdot}, x_{v\cdot 2}) = \sum \exp\left(\sum_i (x_{vi2}\beta_i)\right) \quad . \tag{8.14}$$

Summiert wird hierbei über alle Antwortmuster mit $X_{v\cdot 2} = x_{v\cdot 2}$ und $Z_{v\cdot} = z_{v\cdot}$. Mit Hilfe der neuen Größe $\kappa^{z_{v\cdot}}$ kann die Verteilung von $Z_{v\cdot}$ folgendermaßen dargestellt werden:

$$Pr\left(Z_{v\cdot} = Z_{v\cdot} \mid X_{v2} \in \mathcal{I}\right) = \frac{\kappa(z_{v\cdot}, x_{v\cdot 2}) \exp\left(\lambda z_{v\cdot}\right)}{\sum_{z_{v\cdot}} \left(\kappa(z_{v\cdot}, x_{v\cdot 2}) \exp\left(\lambda z_{v\cdot}\right)\right)} \quad . \tag{8.15}$$

Zunächst soll ein Algorithmus zur Berechnung von $\kappa(z_{v\cdot}, x_{v\cdot 2})$ vorgestellt werden. Dazu werden die folgenden Bezeichnungen benötigt:

$$\xi = x_{v\cdot 2} - x_{v\cdot 1} \tag{8.16}$$

$$\Psi\left(X_{v2}\right) = \exp\left(\sum_i (x_{vi2}\beta_i)\right) \quad . \tag{8.17}$$

Weiterhin bezeichnen wir mit X_{v1}^{-} die Teilmenge der im ersten Untersuchungszeitpunkt falsch beantworteten Items, mit X_{v1}^{+} die Teilmenge der im ersten Untersuchungszeitpunkt richtig beantworteten Items. Mit m^{-}, bzw. m^{+} sei die Zahl der Items dieser Mengen bezeichnet. Startpunkt des Algorithmus ist das Antwortmuster im ersten Messzeitpunkt. Bei diesem besitzt die Testgröße $Z_{v\cdot}$ den Wert m. Aus x_{v1} kann jedes beliebige x_{v2} erzeugt werden, indem man bei jedem Item i mit $x_{vi1} \neq x_{vi2}$ eine Transformation $\Gamma(x_{v1i})$ durchführt, so dass

$$x_{vi2} = \Gamma(x_{v1i}) = 1 - x_{v1i}$$

gilt. Wie man leicht sieht, kann durch solche Transformationen jedes beliebige Antwortmuster x_{v2} aus dem im ersten Messzeitpunkt vorliegenden Muster x_{v1} erzeugt werden.

Um vom Antwortmuster x_{v1} mit Summenscore $x_{v\cdot 1}$ auf ein Muster x_{v2} mit Summenscore $x_{v\cdot 2}$ zu kommen, müssen im Fall $\xi \geq 0$ mindestens $|\xi|$ Transformationen in der Itemmenge X_{v1}^{-} durchgeführt werden, um ein Antwortmuster x_{v2} mit Summenscore $x_{v\cdot 2}$ zu erzeugen. Falls $\xi < 0$ gilt, müssen mindestens ξ Operationen bei Items aus der Itemmenge X_{v1}^{+} durchgeführt werden. Durch jede an einem Item i durchgeführte Transformation Γ verringert sich die Testgröße $Z_{v\cdot}\left(x_{v2}\right)$ eines Antwortmusters x_{v2} um den Wert 1. Maximaler Wert der Testgröße $Z_{v\cdot}\left(x_{v2}\right)$ ist daher $m - |\xi|$.

Um ein Antwortmuster aus $\mathcal{I}$ zu erzeugen, müssen in jedem Fall $|\xi|$ Γ-Transformationen durchgeführt werden. Bei genau $|\xi|$ solcher Transformationen besitzt die Testgröße stets den Wert

$$Z_{v\cdot}\left(x_{v2}\right) = m - |\xi| \quad .$$

Um ein Antwortmuster mit niedrigerer Testgröße zu erzeugen, müssen weitere Γ-Transformationen angewendet werden. Um

$$X_{v\cdot 2} = x_{v\cdot 2}$$

sicherzustellen, müssen stets $2k$ zusätzliche Transformationen gleichzeitig durchgeführt werden: jeweils k Transformationen bei den Items aus X_{v1}^{-}, und k Transformationen bei den Items aus X_{v1}^{+}. Die Antwortmuster der Menge $\mathcal{I}$ können daher mit jeweils $|\xi| + 2k$ Γ-Transformationen erzeugt werden. Dabei gilt

$$\begin{aligned} 0 \leq k \leq \min(m^{+} - |\xi|\,, m^{-}) \qquad & \text{falls } \xi > 0 \\ 0 \leq k \leq \min(m^{+}, m^{-} - |\xi|) \qquad & \text{falls } \xi \leq 0 \quad . \end{aligned}$$

Ein mit insgesamt $|\xi| + 2k$ Transformationen erzeugtes Antwortmuster führt zu der Testgröße $Z_{v\cdot} = m - |\xi| - 2k$.

Um $\kappa(z_{v\cdot}, x_{v\cdot 2})$ zu berechnen, untersuchen wir zunächst ein Antwortmuster x_{v2}, das sich von x_{v1} nur bei einem Item i^{*} unterscheidet. Es ergibt sich

$$\begin{aligned} \Psi\left(X_{v2}\right) &= \exp\left(\sum_{i}\left(x_{vi2}\beta_i\right)\right) * \frac{\exp\left(-x_{vi^{*}2}\beta_{i^{*}}\right)}{\exp\left(-x_{vi^{*}2}\beta_{i^{*}}\right)} \\ &= \begin{cases} \Psi\left(X_{v1}\right)\exp\left(\beta_{i^{*}}\right); & \text{falls } i^{*} \in X_{v1}^{+} \\ \Psi\left(X_{v1}\right)\exp\left(-\beta_{i^{*}}\right); & \text{falls } i^{*} \in X_{v1}^{-} \end{cases} \quad . \end{aligned} \tag{8.18}$$

Daher erhält man

$$\kappa(m-1, x_{v\cdot 1} - r) = \Psi\left(X_{v1}\right)\left(\sum_{i \in X_{v1}^{+}} \exp\beta_i + \sum_{i \in X_{v1}^{-}} \exp(-\beta_i)\right) \quad . \tag{8.19}$$

Hierbei ist $r = x_{v\cdot 1} + 1$, falls $i^{*} \in X_{v1}^{+}$, und $r = x_{v\cdot 1} - 1$ andernfalls. $\kappa(m-1, r)$ betrifft zwar keines der Antwortmuster in $\mathcal{I}$, eignet sich aber hervorragend, um das Berechnungsprinzip für $\kappa^{z_{v\cdot}}$ darzustellen.

Als nächsten Schritt untersuchen wir ein Antwortmuster x_{v2}, das sich in zwei Items j und l von x_{v2} unterscheidet, wobei $j \in X_{v1}^{+}$ und $l \in X_{v1}^{-}$. Dann erhalten wir

$$\Psi\left(X_{v2}\right) = \Psi\left(X_{v1}\right)\exp\left(\beta_j - \beta_l\right)$$

sowie

$$\begin{aligned} \kappa(m-2, x_{v\cdot 1}) &= \Psi\left(X_{v1}\right)\sum_{i \in X_{v1}^{+}} \exp\beta_i \sum_{j \in X_{v1}^{-}} \exp\left(-\beta_j\right) \\ &= \Psi\left(X_{v1}\right)\gamma^{1}\left(X_{v1}^{+}\right)\gamma^{1}\left(X_{v1}^{-}\right) \quad . \end{aligned} \tag{8.20}$$

Hierbei ist $\gamma^1\left(X_{v1}^+\right)$ die elementare symmetrische Funktion

$$\gamma^1\left(X_{v1}^+\right) = \sum_{i \in X_{v1}^+} exp\left(\beta_i\right) \quad . \tag{8.21}$$

Äquivalent wird $\gamma^1\left(X_{v1}^-\right)$ definiert. Wie man leicht sieht, kann man die Berechnung von $\kappa(m-r-s, x_{v\cdot 1}+r-s)$ für r Transformationen innerhalb von X_{v1}^- und s Transformationen innerhalb von X_{v1}^+ ebenfalls auf elementare symmetrische Funktionen zurückführen. Somit können auch die κ-Terme für die Antwortmuster aus $\mathcal{I}$ mit Hilfe elementarer symmetrischer Funktionen berechnet werden:

$$\kappa\left(m-|\xi|-2k, x_{v\cdot 2}\right) = \begin{cases} \Psi\left(X_{v1}\right)\gamma^{\xi+k}\left(X_{v1}^+\right)\gamma^k\left(X_{v1}^-\right) & \text{falls} \quad \xi > 0 \\ \Psi\left(X_{v1}\right)\gamma^{k}\left(X_{v1}^+\right)\gamma^{\xi+k}\left(X_{v1}^-\right) & \text{falls} \quad \xi \leq 0 \end{cases} \quad . \tag{8.22}$$

Nach Berechnung der κ-Terme kann die Verteilung aus (8.15) leicht berechnet werden.

Kapitel 9

Eigenschaften der vorgestellten Methoden

Die Eigenschaften der optimalen Personenfittests für Veränderung werden zum einen über die Teststärke untersucht, zum anderen mit Hilfe eines speziell konstruierten Testdatensatzes, der verschiedene Arten von Abweichungen aufweist. Um das Verständnis des nächsten Abschnitts zu erleichtern, wollen wir an dieser Stelle einige Begriffe erläutern: Als *Zahl eingehender Items* bezeichnen wir die Zahl m von Items, die zu Zeitpunkt 2 anders beantwortet werden als zu Zeitpunkt 1. Als *Abweichungsgruppe* bezeichnen wir eine Gruppe simulierter Antwortmuster, bei denen die Veränderung auf gleiche Art erzeugt wurde, d.h. eine Gruppe von Antwortmustern mit gleicher Art von Fehlspezifikation der Veränderung. Die uns interessierende Größe ist dabei die relative Ablehnhäufigkeit in einer Personengruppe mit gleicher Verteilung der Testgröße.

9.1 Untersuchte Problemstellungen

Der Testdatensatz besteht aus einer Gruppe Versuchspersonen, bei denen die Veränderung gemäß eines bestimmten LLTMs erfolgt, und 5 Personengruppen, bei denen dies nicht der Fall ist. Es wird folgendes (einfaches) LLTM für einen einfachen 2×2-Versuchsplan mit 2 Zeitpunkten und 2 Untersuchungsgruppen (Versuchs- und Kontrollgruppe) angenommen:

$$P(X_{vgti} = x_{vgti}) = \frac{exp\left[x_{vgti}(\theta_v - \beta_i + \delta_{gt})\right]}{1 + exp\left[\theta_v - \beta_i + \delta_{gt}\right]} \quad . \tag{9.1}$$

Dabei steht θ_v für den Fähigkeitsparameter der Person v, β_i für die Schwierigkeit eines Items i sowie δ_{gt} für die Veränderung in Untersuchungsgruppe g im Zeitpunkt t.

Es wird ein Datensatz mit insgesamt 3000 Personen künstlich erzeugt, der sich aus 6 Abweichungsgruppen mit jeweils 500 Personen zusammensetzt. In jeder Abweichungsgruppe

werden 250 Personen gemäß den Parameterwerten der Kontroll- resp. Versuchsgruppe erzeugt. Insgesamt lassen sich die Abweichungsgruppen folgendermaßen beschreiben:

- Gruppe 1 enthält die VPs, bei denen die Daten rein mit obigem LLTM erzeugt wurden. Für die Versuchsgruppe wird dabei $\delta_{12} = 0.2$, für die Kontrollgruppe $\delta_{22} = -2$ angenommen.

- In den Gruppen 2 bis 4 wird zu der durch das LLTM gegebenen Veränderungen ein Störterm ξ dazuaddiert. Also beträgt der tatsächliche Veränderungswert einer Person:
$$\delta_{gtv}^{akt} = \delta_{gt} + \xi_v \ . \tag{9.2}$$
Der Störterm ξ_v hängt immer mit der tatsächlichen Fähigkeit zusammen: In den Gruppen 2 und 3 wird der Zusammenhang zwischen Störterm ξ_v und Fähigkeit θ_v bei einer Person v durch eine exakte lineare Beziehung der Form
$$\xi_v = b * \theta_v \tag{9.3}$$
modelliert. In Gruppe 2 wird $b = 0.8$, in Gruppe 3 $b = -0.8$ angenommen. Die Gruppe 4 verwendet eine quadratische Funktion als Zusammenhang zwischen Fähigkeit und Störung, nämlich
$$\xi_v = 0.1\theta_v^2 \ . \tag{9.4}$$

- In Gruppe 5 wird versucht, Bias bezüglich der Veränderung zu modellieren. Dabei wird zu 10 (von insgesamt 30) Items verschiedener Schwierigkeitsgrade 1.5 (bzw. -1.5) zu der durch das LLTM gegebenen Veränderung dazuaddiert.

- In Gruppe 6 wird schließlich untersucht, ob Personenfittests für Veränderung auch andere Formen von Fehlspezifikation entdecken können, z.B. eine durch eine falsch angenommene Traitfunktion hervorgerufene Fehlspezifikation. Dabei wird von linearen Traitfunktionen mit unterschiedlichen Trennschärfeparametern ξ_i ausgegangen:
$$P(X_{vit} = 1) = max(min(1, \xi_i * (\theta_v - \beta_i + \delta_{gt})), 0) \ . \tag{9.5}$$
Als Fähigkeits- und Schwierigkeitsparameter dieses letzten Modells werden jeweils die gleichen Werte wie in Gruppe 1 verwendet. Als Trennschärfeparameter wird für ein Item i der Wert $\xi_i = 1 + 0.5 \times (-1^i)$ verwendet, d.h. je nach Index des Items die Trennschärfe 0.5 bzw. 1.5.

Jeweils 10 Personen in einer Abweichungsgruppe (d.h. jeweils 5 Personen gleicher Abweichungsform in Kontroll- resp. Versuchsgruppe) besitzen gleiche Fähigkeitswerte. Als Fähigkeitswerte werden $-2.5, -2.4, -2.3, ..., 0, 0.1, ..., 2.5$ verwendet. Als Schwierigkeitsparameter werden -1.5, -1.4,...,0, 0.1,..., 1.5 verwendet. Damit ist sowohl für die Schwierigkeits- als auch für die Fähigkeitsparameter eine große Spannbreite bezüglich der verwendeten Werte gewährleistet.

9.2 Technische Einzelheiten der Simulation

Die Antworten der Versuchspersonen wurden folgendermaßen bestimmt: Es wurde für jedes Item und jede Versuchsperson eine tatsächliche Veränderung δ_v^{akt} berechnet. Die Schwierigkeiten β_{i2} der Items im zweiten Zeitpunkt wurden dann mittels

$$\beta_{i2} = \beta_{i1} + \delta_{gt} + \xi_v = \beta_{i1} + \delta_v^{akt} \tag{9.6}$$

bestimmt. Gemäß dem zugrunde gelegten Traitmodell wurde dann die Lösungswahrscheinlichkeit für Item i und Person v berechnet, und zwar mittels

$$P(X_{vti} = 1) = \frac{exp(\theta_v - \beta_{ti})}{1 + exp(\theta_v - \beta_{ti})} \tag{9.7}$$

für die Gruppen 1-5 und eines nichtlogistischen Traitmodells für die Gruppe 6. Mit einem für jede Abweichungsgruppe spezifischen Startwert wurde dann eine Zufallszahl zwischen 0 und 1 erzeugt. Falls die Zufallszahl kleiner als die berechnete Lösungswahrscheinlichkeit war, wurde Item i als von Person v gelöst angenommen, andernfalls als nicht gelöst.

9.3 Untersuchungsziele

Mit diesem Testdatensatz sollen folgende Problemstellungen untersucht werden:

1. Besitzt die Annahme eines bestimmten Zusammenhangs zwischen Fähigkeitswert und Veränderungsparameter einen Einfluss auf die Teststärke der Personenfittests?
2. Ist die Teststärke der Personenfittests für jede Fähigkeitsstufe gleich?
3. Sind die Personenfittests robust gegenüber Fehlannahmen der zugrunde liegenden Traitfunktion?

Natürlich kann mittels dieses Testdatensatzes keine endgültige Antwort auf die gestellten Fragen gegeben werden. Die Ergebnisse der Untersuchung des Testdatensatzes sind daher als Ergänzung der in Kapitel 10 folgenden Teststärkeuntersuchungen anzusehen. Alle Ergebnisse resultieren aus Tests mit einer punktförmigen Nullhypothese bei Signifikanzniveau $\alpha = 0.05$.

9.4 Auszählung der relativen Häufigkeit von Ablehnungen pro Abweichungsgruppe

Auf die nach obiger Beschreibung erzeugten Datensätze wurden Tests mit 2 verschiedenen Nullhypothesen angewendet:

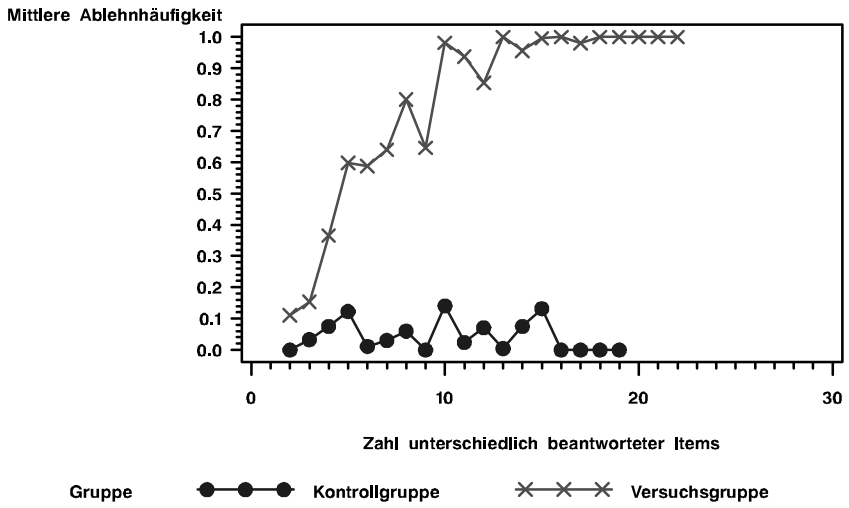

Abbildung 9.1: relative Ablehnhäufigkeit vs. Zahl eingehender Items getrennt nach Zugehörigkeit zu Untersuchungsgruppe resp. Kontrollgruppe. Veränderung unabhängig von Fähigkeit.

- H_0: $\delta_{gt} = 0.2$ gegen H_A: $\delta_{gt} \neq 0.2$
- H_0: $|\delta_{gt} - 0.2| \leq 0.1$ gegen H_A: $|\delta_{gt} - 0.2| \geq 0.1$.

Für jede Versuchsperson konnte auf diese Weise eine Ablehnhäufigkeit bestimmt werden, die die folgenden Werte annimmt:

- 0 falls die Testgröße einer Versuchsperson in den Annahmebereich fällt,
- 1 falls die Testgröße in das Innere des Ablehnbereichs fällt, oder
- γ_1 resp. γ_2 falls die Testgröße die Grenzen C_1 resp. C_2 des Ablehnbereichs trifft.

In einem ersten Schritt wurde für die (aus Abweichungsgruppe, Zugehörigkeit zu Versuchs- resp. Kontrollgruppe und der Zahl der in den Test eingehenden Items entstehende) Klassifikation der Versuchspersonen Summe und arithmetisches Mittel dieser Ablehnhäufigkeit berechnet. Die Ergebnisse sind in Grafik 9.1 dargestellt. Da die Nullhypothese den vorausgesetzten Parameterwerten der Kontrollgruppe entspricht, sollte in der Kontrollgruppe von Abweichungsgruppe 1 die Ablehnhäufigkeit kleinere Werte als $\alpha = 0.05$ annehmen. Wie man sieht, ist dies in der Regel der Fall.

Die Versuchsgruppe, für die ein Veränderungsparameter von 2 vorausgesetzt wurde, sollte durchgehend hohe Werte für die relative Ablehnhäufigkeit aufweisen. Dies ist auch der Fall. Wie man weiterhin sieht, steigt die Ablehnhäufigkeit mit der Zahl beteiligter Items stark an. Wie zu erwarten, decken die Tests umso häufiger Fehlspezifikationen auf, je mehr Items in den Test eingehen. Dieses Verhalten hat Auswirkungen auf die Trennschärfe der Tests in Extrembereichen: Bei extremen (d.h. sehr hohen oder sehr niedrigen) Fähigkeitswerten

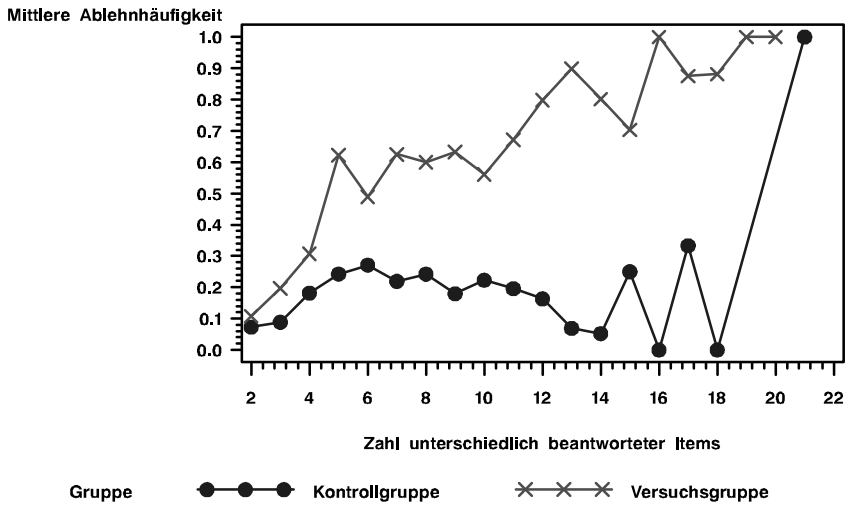

Abbildung 9.2: relative Ablehnhäufigkeit vs. Zahl eingehender Items getrennt nach Zugehörigkeit zu Untersuchungsgruppe resp. Kontrollgruppe. Positiver linearer Zusammenhang zwischen Veränderung und Fähigkeit.

wächst (resp. fällt) die Lösungswahrscheinlichkeit eines Items i zwischen Zeitpunkt 1 und Zeitpunkt 2 nur unwesentlich, da dann immer

$$Pr\left(X_{vi1} = 1 \,|\theta_v\right) \approx Pr\left(X_{vi1} = 1 \,|\theta_v + \delta_{g2}\right) \tag{9.8}$$

gilt. Daraus folgt aber eine niedrige Zahl von Items, die an den beiden Testzeitpunkten unterschiedlich gelöst werden. Abb. A.1 (vgl. Anhang) zeigt den Zusammenhang zwischen der Ablehnungshäufigkeit und der Fähigkeit. Der oben beschriebene Effekt ist in der Versuchgruppe deutlich erkennbar.

Die Abweichungsgruppen 2 und 3 weisen eine Verzerrung auf, die linear mit dem Fähigkeitsparameter zusammenhängt. In Abweichungsgruppe 2 ist dies eine steigende, in Abweichungsgruppe 3 eine fallende lineare Funktion. Die Verzerrung überlagert hierbei additiv die „wahre“ Veränderung ohne Verzerrung. Die steigende lineare Funktion bewirkt bei Personen mit hoher Fähigkeit eine größere Fähigkeitssteigerung als bei Personen mit niedriger Fähigkeit. Andererseits führt aber eine gegebene Änderung δ_0 zu einer umso geringeren Änderung der Lösungswahrscheinlichkeit $P(X_{vit} = 1)$ eines Items mit Schwierigkeit β_i, je höher der Fähigkeitswert θ_v einer Person ist, wobei immer $\theta_v > \beta_i$ vorausgesetzt werden muss. Daher ähnelt ein Antwortmuster einer Person mit hohen Fähigkeitswerten bei starker Verzerrung sehr dem Antwortmuster, das ohne Verzerrung entstanden wäre, falls diese Verzerrung eine Fähigkeitssteigerung bewirken würde. Dieser Effekt ist deutlich in Abbildung A.2 zu erkennen: Bei sehr hohen Fähigkeiten addiert sich eine zur Fähigkeit gleichgerichtete Verzerrung zu einer zur Fähigkeit entgegengerichteten Veränderung, so dass sich Verzerrung und Veränderung gegenseitig auslöschen. In der Kontrollgruppe hingegen findet man bei extremen Fähigkeiten eine leicht erhöhte Ablehnhäufigkeit.

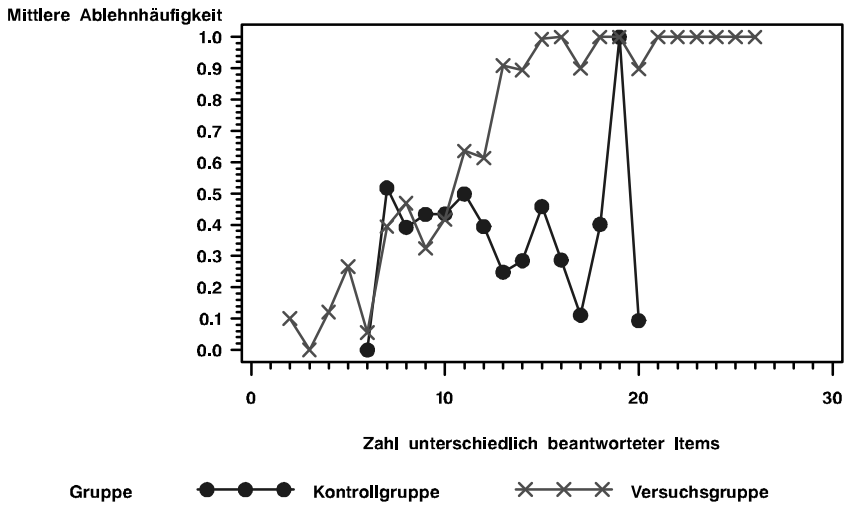

Abbildung 9.3: relative Ablehnhäufigkeit vs. Zahl eingehender Items getrennt nach Zugehörigkeit zu Untersuchungsgruppe resp. Kontrollgruppe. Negativer linearer Zusammenhang zwischen Veränderung und Fähigkeit.

An Abbildung 9.2 kann man ablesen, dass in der Kontrollgruppe die Ablehnhäufigkeit zwar durchgängig höher als in Abweichungsgruppe 1 ist, aber insgesamt auf recht niedrigem Niveau. In der Versuchsgruppe zeigen sich in Abweichungsgruppe 2 durchgängig niedrigere relative Ablehnhäufigkeiten als in Abweichungsgruppe 1.

Abb. A.2 zeigt den Zusammenhang zwischen Fähigkeit und Ablehnhäufigkeit bei einer negativen Korrelation zwischen Fähigkeit und Veränderung. Aufgrund des der Fähigkeit entgegengesetzten Veränderungsparameters in der Versuchsgruppe ist ein Auslöschungseffekt bei extrem niedrigen Fähigkeiten zu bemerken. Bei sehr hohen Fähigkeiten ist ein Überlagerungseffekt zu beobachten: Veränderung und Verzerrung verstärken sich gegenseitig. In der Kontrollgruppe ergeben sich hohe Ablehnhäufigkeiten bei extremen Fähigkeiten, sowie sehr niedrige Ablehnhäufigkeiten bei mittleren Fähigkeiten.

In Abb. 9.3 zeigen sich eher niedrige relative Ablehnhäufigkeiten in der Kontrollgruppe, sowie in der Experimentalgruppe niedrigere relative Ablehnhäufigkeiten als in der Abweichungsgruppe 1. Ferner zeigt sich – wie auch in Abweichungsgruppe 2 – eine relativ große Variabilität der Ablehnhäufigkeiten: diese steigen nicht so klar mit wachsender Zahl eingehender Items an wie in Abweichungsgruppe 1, und weisen größere Schwankungen auf.

Ein interessantes Ergebnis liegt in der Abweichungsgruppe mit quadratischer Verzerrung vor (vgl. auch Abb. 9.4). In der Versuchsgruppe steigt die relative Ablehnhäufigkeit schneller an als in der Versuchsgruppe ohne Verzerrung. In der Kontrollgruppe hingegen schwankt die relative Ablehnhäufigkeit um 0.05, d.h. die Verzerrung der Kontrollgruppe wird sehr selten erkannt. A.4 zeigt den Zusammenhang zwischen Fähigkeit und Ab-

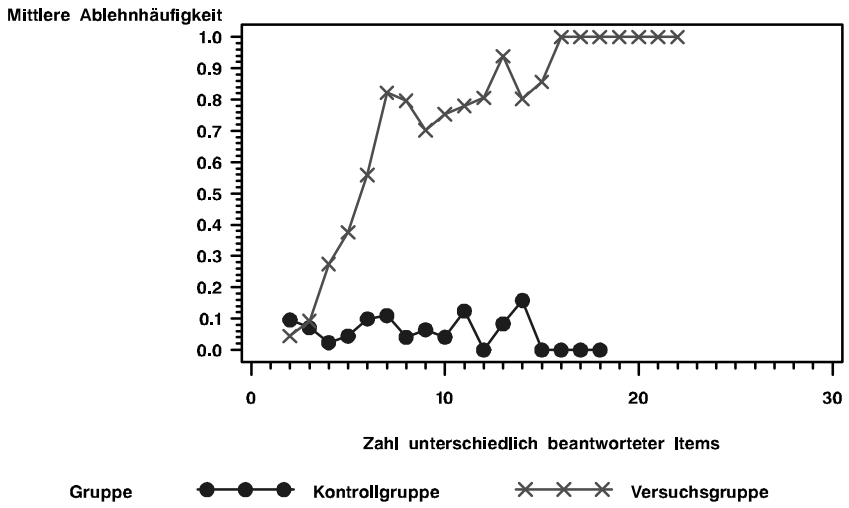

Abbildung 9.4: Relative Ablehnhäufigkeit vs. Zahl eingehender Items getrennt nach Zugehörigkeit zu Untersuchungsgruppe resp. Kontrollgruppe. Quadratischer Zusammenhang zwischen Veränderung und Fähigkeit.

lehnhäufigkeit bei quadratischem Zusammenhang zwischen Fähigkeit und Veränderung. Das Bild ähnelt sehr Abb. A.1, weist in der Versuchsgruppe jedoch einen stärkeren Abfall der Ablehnhäufigkeit bei extremen Fähigkeiten auf.

Ein ähnliches Bild zeigt sich bei der Untersuchung der Abweichungsgruppe „Veränderungs-Bias in 10 Items“ (vgl. auch Abb. 9.5): gutes Differenzierungsvermögen in der Versuchsgruppe, aber kein Erkennen von Verzerrungen in der Kontrollgruppe. Dies kann darauf zurückzuführen sein, dass der Veränderungs-Bias hier symmetrisch konstruiert wurde: Bei einer Hälfte der Items mit Veränderungs-Bias im zweiten Messzeitpunkt wurde die Itemschwierigkeit um -1.5 verändert, bei der anderen Hälfte um $+1.5$. Verglichen mit dem unverzerrten Modell werden einige Items zu häufig beantwortet, andere dagegen zu selten. Dies führt dazu, dass die relative Häufigkeit der Items, die im zweiten, nicht aber im ersten Messzeitpunkt richtig beantwortet wurden, ungefähr gleich bleibt.

Bei der Untersuchung von Fähigkeit und Ablehnhäufigkeit (Abb. A.5) zeigt sich im Vergleich zu Abb. A.1 eine niedrigere Ablehnhäufigkeit bei niedrigen Fähigkeiten sowie eine erhöhte Ablehnhäufigkeit bei hohen Fähigkeiten in der Versuchsgruppe. In der Kontrollgruppe ergibt sich durchgängig eine niedrigere Ablehnhäufigkeit als in Abb. A.1.

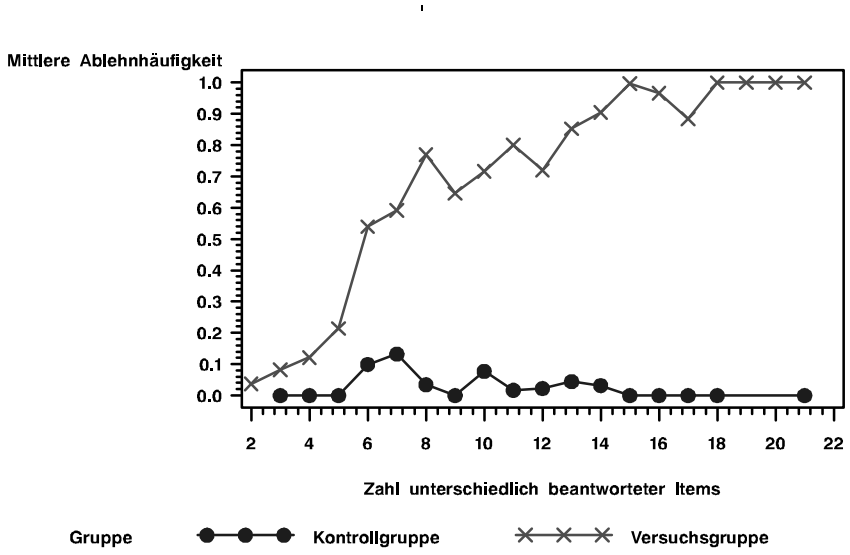

Abbildung 9.5: Relative Ablehnhäufigkeit vs. Zahl eingehender Items getrennt nach Zugehörigkeit zu Untersuchungsgruppe resp. Kontrollgruppe. Verzerrung durch Differential Item Functioning in 10 Items.

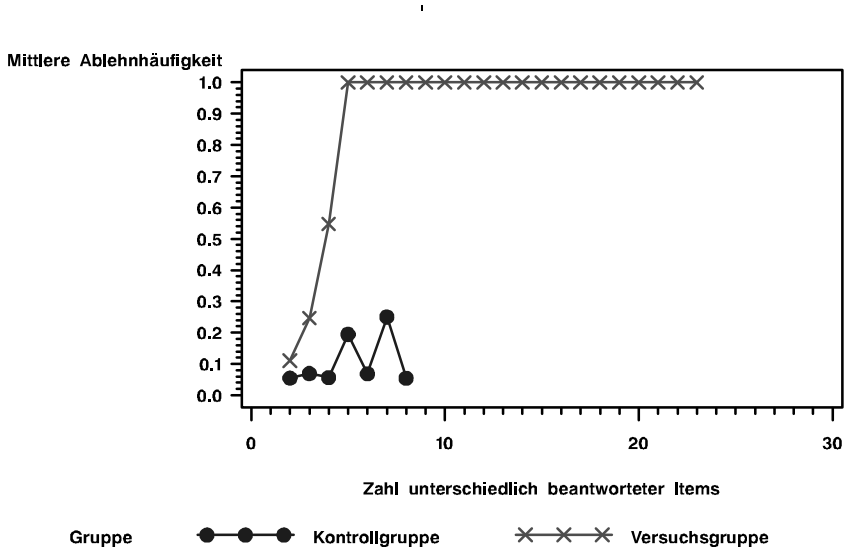

Abbildung 9.6: Relative Ablehnhäufigkeit vs. Zahl eingehender Items getrennt nach Zugehörigkeit zu Untersuchungsgruppe resp. Kontrollgruppe. Verzerrung durch lineare Traitfunktionen mit unterschiedlichen Fähigkeiten.

Schließlich soll auch das Ergebnis aus Abweichungsgruppe 6 dargestellt werden, wo die Veränderungsparameter verglichen mit Abweichungsgruppe 1 unverändert geblieben sind. Die Ablehnhäufigkeiten sind in Abb. 9.6 dargestellt. In der Kontrollgruppe ergeben sich, im Vergleich zu allen anderen Abweichungsgruppen, die höchsten Ablehnhäufigkeiten. In der Versuchsgruppe werden höhere Ablehnhäufigkeiten als in der Kontrollgruppe beobachtet, wobei die Ablehnhäufigkeit mit der Zahl eingehender Items zunimmt. Abb. A.6 zeigt den Zusammenhang zwischen Fähigkeit und Ablehnhäufigkeit. In der Kontrollgruppe ist eine äußerst starke Schwankung der Ablehnhäufigkeit zu bemerken, in der Versuchsgruppe dagegen liegt die relative Ablehnhäufigkeit bei mittleren und hohen Fähigkeiten stets bei 1.

Weiterhin wurden Personenfittests zu einer intervallförmigen Nullhypothese durchgeführt, nämlich H_0: $0.1 \leq \delta \leq 0.3$ gegen H_A: $\delta \leq 0.1$ oder $\delta \geq 0.3$. Interessanterweise ergeben sich hier (fast) die gleichen relativen Ablehnhäufigkeiten wie bei den punktförmigen Nullhypothesen (vgl. auch Abbildungen in Anhang A). Vergleicht man die Annahmebereiche der Tests mit punktförmiger Nullhypothese mit den Annahmebereichen der Tests mit intervallförmiger Nullhypothese, so ergeben sich kaum Unterschiede in den Ablehnhäufigkeiten. In der Simulation besitzt die Wahl der Hypothesenarten daher nur einen geringen Einfluss, wie man leicht durch Vergleich der Grafiken 9.1 bis 9.6 mit den korrespondierenden Grafiken A.7 bis A.12 erkennen kann.

9.5 Diskussion

Mit dem hier vorgestellten, durch Simulation erzeugten Testdatensatz wurde versucht, explorativ einige Eigenschaften von Personenfittests für die Veränderungsmessung aufzudecken. Im Rahmen der Diskussion der Simulationsergebnisse wollen wir folgende Punkte unterscheiden:

- Einfluss der Art der H_0: intervallförmige H_0 vs. punktförmige H_0
- Robustheit der optimalen Personenfittests
- Aufdecken von Differential Item Functioning
- Globaler Einfluss von Zahl eingehender Items, Fähigkeit und tatsächlicher Veränderung

Einfluss der Art der H_0: intervallförmige H_0 vs. punktförmige H_0

In der Simulation wurde untersucht, ob die Wahl einer intervallförmigen Nullhypothese $\delta \in [0.2; 0.1]$ zu einem anderen Verhalten der Personenfittests führt als die Verwendung der punktförmigen Nullhypothese $\delta = 0.2$. Die Unterschiede hinsichtlich der Ablehnhäufigkeit

erwiesen sich für die gewählten Nullhypothesen als überraschend gering.

Andererseits bieten intervallförmige Nullhypothesen interpretative Vorteile, da man in das H_0-Intervall Information über die praktische Signifikanz eines Effektes einfließen lassen kann. Wenn viele Items in die Personenfittests eingehen, vermeidet man mit intervallförmigen Nullhypothesen, dass wegen ihrer geringen Größe nicht interessierende Effekte als signifikant ausgewiesen werden. Nach Meinung des Autors empfiehlt sich bei der Untersuchung des Personenfits diese Art der Nullhypothese, da sie eine genauere Spezifikation des zu untersuchenden Gegenstandes ermöglicht. Da bei Personenfituntersuchungen meist eine große Zahl von Tests gleichzeitig durchgeführt werden, ist hier eine praktisch möglichst relevante Definition der Nullhypothese angebracht. Somit ist dieses Vorgehen eine interessante Alternative zum „klassischen“ Konzept der Planung der Stichprobengröße (vgl. dazu: [Cohen, 1969]).

Insgesamt gesehen spielt diese „praktisch relevante Signifikanz“ hauptsächlich bei Untersuchungen mit vielen Items eine Rolle. Bei einer geringen Zahl eingehender Items m ergeben sich, wie man aus Abb. A.7 bis Abb. A.12 sehen kann, nur geringfügige Unterschiede zu den üblichen Tests mit punktförmiger Nullhypothese.

Robustheit der optimalen Personenfittests

Der Simulationsansatz ermöglicht Aussagen über die Robustheit der optimalen Personenfittests, wenn eine zusätzliche Verzerrung die Daten beeinflusst. Dazu verwenden wir die Ergebnisse für die Versuchsgruppe. In dieser wurde eine vorgegebene Veränderung von $\delta = -2$ mit einer Verzerrung kombiniert. Es zeigte sich, dass bei allen Verzerrungen die in der Experimentalgruppe auftretende Abweichung von der Nullhypothese mit hoher Wahrscheinlichkeit entdeckt werden konnte, wenn sich Verzerrung und Veränderung nicht gegenseitig auslöschten. Die relative Ablehnhäufigkeit steigt dabei mit der Zahl eingehender Items an.

Bei Abweichungsgruppe 5 (= Bias bezüglich der Veränderung) scheint sich das Verhalten der Tests gegenüber der Situation ohne Verzerrung sogar zu verbessern: Die relative Ablehnhäufigkeit in der Kontrollgruppe ist – bei allen Fähigkeitsstufen – eher niedriger als beim Teildatensatz ohne Verzerrung.

Gewisse Anzeichen für robustes Verhalten der Personenfittests sind auch bei den Versuchspersonen mit linearer Itemantwortfunktion und variierenden Trennschärfen zu erkennen. Hier ist in der Versuchsgruppe die Ablehnhäufigkeit höher als in der Versuchsgruppe der Personen ohne Verzerrung. Andererseits ist in der Kontrollgruppe eine sehr große Schwankungsbreite zu erkennen. Verbesserten Eigenschaften der Personenfittests in der Versuchsgruppe steht also ein uneindeutiges Verhalten in der Kontrollgruppe gegenüber.

Aufdecken von Bias bezüglich der Veränderung

Die künstlich erzeugten Daten mit Bias bezüglich der Veränderung weisen aufgrund der Konstruktion der Stichprobe keinen Testbias auf: Bei 5 Items wurde die Veränderung um 1.5 erhöht, bei 5 anderen Items um 1.5 erniedrigt. Gemäß [Roznowski and Reith, 1999] gleichen sich die Verzerrungen der einzelnen Items aus, so dass insgesamt kein Testbias vorliegt.

Die optimalen Tests auf Personenfit können diese Verzerrung nicht erkennen. In der Kontrollgruppe scheint die relative Ablehnhäufigkeit sogar noch etwas niedriger zu sein als im unverzerrten Teildatensatz. In der Versuchsgruppe entspricht die relative Ablehnhäufigkeit in etwa der Ablehnhäufigkeit des unverzerrten Datensatzes.

Dieses Ergebnis widerspricht somit – bei der gewählten allgemeinen Alternativhypothese – der Aussage von [Ponocny, 2000], wo die Verwendung von Personenfittests für Untersuchungen auf Bias bezüglich der Veränderung vorgeschlagen wird. Dies mag allerdings darauf zurückzuführen sein, dass keine Alternativhypothese gewählt wurde, die speziell für das Auftreten von Bias bezüglich der Veränderung geeignet ist. Ein weiterer Grund für die fehlende Empfindlichkeit gegenüber Bias bezüglich der Veränderung mag in dem fehlenden Testbias liegen: Sei dazu X_{01} die Zahl der verzerrten Items, die nur im zweiten Zeitpunkt gelöst wurden, sowie X_{10} die Zahl der verzerrten Items, die nur im ersten Zeitpunkt gelöst wurden. Aufgrund des symmetrischen Aufbaus des Testdatensatzes ist

$$EX_{10} \approx EX_{01} \quad .$$

Die konstruierte Verzerrung sorgt also zumindest im Mittel nicht für ein Abgleiten der Testgröße in den Ablehnbereich.

Globaler Einfluss von Zahl eingehender Items, Fähigkeit und tatsächlicher Veränderung

In der Experimentalgruppe zeigt sich stets ein Zusammenhang der Ablehnhäufigkeit mit der Zahl eingehender Items m. Dagegen ist das Verhalten in der Kontrollgruppe uneinheitlich. Fähigkeit und aktuelle Veränderung zeigen sich meist als wichtige Einflussfaktoren. Dabei nimmt die relative Ablehnhäufigkeit in der Experimentalgruppe mit wachsender tatsächlicher Veränderung ab. Da $\delta = -2$ als unverzerrte Veränderung vorgegeben wurde, bedeutet ein Ansteigen der tatsächlichen Veränderung immer eine Verkleinerung des Effekts.

Sachlich bedeutsamer ist der negative Schwierigkeitsparameter für die Fähigkeit in der Experimentalgruppe. Bei extremen Fähigkeiten wird eine aberrante Veränderung seltener entdeckt. Dies gilt für alle Abweichungsgruppen außer Gruppe 6. Dieses Verhalten lässt

sich mit der Konstruktion des Testdatensatzes erklären: In den Extrembereichen der logistischen Itemantwortfunktion bewirkt eine Veränderung des Schwierigkeitsparameters eine geringere Änderung der Lösungswahrscheinlichkeit eines Items als im mittleren Schwierigkeitsbereich. Daraus folgt aber, dass die Testgröße der untersuchten Tests meist im Annahmebereich bleibt.

Kapitel 10

Teststärkenuntersuchungen

Wichtig für die Verwendung optimaler Personenfittests ist das Wissen über die Teststärke dieser Tests gegenüber bestimmten Alternativen. Ebenso wie im Kapitel 9 wollen wir hierbei praxisrelevante Alternativen voraussetzen, die teilweise der Nullhypothese widersprechen. So werden wir folgende Fälle untersuchen:

- Teststärke bei „normalen“ Abweichungen von der Nullhypothese
- Teststärke bei extremen Fähigkeitswerten
- Teststärke bei Zusammenhang zwischen Veränderung und Fähigkeit im ersten Messzeitpunkt
- Vergleich der Teststärken bei intervallförmigen und punktförmigen Nullhypothesen.

In diesem Kapitel werden wir folgende Bezeichnungen verwenden:

$$\begin{aligned} p_0 &= \frac{\exp(\delta_0)}{1+\exp(\delta_0)} \\ p_{akt} &= \frac{\exp(\delta_{akt})}{1+\exp(\delta_{akt})} \quad . \end{aligned}$$

δ_0 ist dabei der Wert für δ_{gt} in der Nullhypothese, δ_{akt} der für eine Person in Wirklichkeit aufgetretene δ_{gt}-Wert. Die Teststärke ergibt sich (bei einer punktförmigen Nullhypothese) durch die folgende Gleichung:

$$\begin{aligned} Pr(X_{v\cdot 2} \in K | \delta = \delta_{akt}) = & \sum_{X_{v\cdot 2} < c_1} P(X_{v\cdot 2} = x_{v\cdot 2} | \delta = \delta_{akt}) + \\ & + \sum_{X_{v\cdot 2} > c_2} P(X_{v\cdot 2} = x_{v\cdot 2} | \delta = \delta_{akt}) + \sum_{i=1}^{2} P(X_{v\cdot 2} = c_i | \delta = \delta_{akt}) \gamma_i \quad . \end{aligned} \tag{10.1}$$

Zur Darstellung der Teststärke benötigen wir eine von der Stichprobengröße unabhängige Größe für die Messung der Effektstärke. Da die Testgröße X_{v2+} parametrisch binomialverteilt mit „Linkfunktion“

$$p(\delta_{gt}) = \frac{exp(\delta_{gt})}{1 + exp(\delta_{gt})} \tag{10.2}$$

ist, können wir die gesuchte Effektstärke auf die Effektstärke für die Untersuchung von Binomialverteilungen zurückführen. Gemäß [Cohen, 1969], bzw. [Cohen, 1992] benötigen wir dazu die Arcussinus-Transformation $arcsin(p(\delta))$ der Wahrscheinlichkeit $p(\delta_{gt})$:

$$AS(p(\delta_{gt})) = arcsin(\sqrt{p(\delta_{gt})} = arcsin(\sqrt{\frac{exp(\delta_{gt})}{1 + exp(\delta_{gt})}} \quad . \tag{10.3}$$

[Cohen, 1969], bzw. [Cohen, 1992] verwenden diese Transformation für die Berechnung der Effektgröße bei nichtrandomisierten Tests für die Binomialverteilung. Wie man sich leicht überlegen kann, gilt die geforderte Translationsinvarianz jedoch auch bei randomisierten Tests. Als Effektgröße ES bei vorgegebenen Parameterwerten δ_0 für die Nullhypothese und δ_{akt} für den tatsächlich aufgetretenen Parameterwert erhält man:

$$ES(\delta_0, \delta_{akt}) = |arcsin(p(\delta_{akt})) - arcsin(p(\delta_0))| \quad . \tag{10.4}$$

10.1 Teststärke in Abhängigkeit von tatsächlicher Veränderung und Zahl eingehender Items

Hierzu berechnen wir zunächst die oben erwähnte Effektstärke bei 5 bis 100 vorgegebenen Items. Die vorgegebenen Werte der Effektstärke liegen zwischen 0 und 0.76. Für eine Fehlerwahrscheinlichkeit erster Art von $\alpha = 0.05$ ist die Teststärke in Abbildung 10.1 zu sehen. Diese Abbildung ist – ebenso wie die Abbildungen 10.2 und 10.3 – als Contourplot zu verstehen: Dargestellt werden Höhenlinien der Teststärke in Abhängigkeit von Effektgröße und der Zahl der eingehenden Items.

Man erkennt sofort die Abhängigkeit zwischen der Zahl der Items und der Teststärke. Dieser Zusammenhang ist jedoch problematisch: Die Zahl eingehender Items m ist definiert als die Zahl der Items, die in beiden Messzeitpunkten unterschiedlich beantwortet wurden, und stellt somit eine Funktion der Scoresummen $X_{vi\cdot}$ pro Item i und Person v über die beiden Untersuchungszeitpunkte dar. Weiterhin bauen die optimalen Tests auf der Verteilung der Testgröße $X_{v\cdot 2}$ auf, unter der Bedingung, dass die Werte für $X_{vi\cdot}$ bekannt sind. Daher hängen die Testgüte und die Testgrößenverteilung von den gleichen Statistiken ab.

Falls viele Itemscores den Wert 0 oder 2 aufweisen, führt dies zu einer niedrigeren Teststärke. Dies bedeutet: In den Extrembereichen besitzen die optimalen Tests auf Personenfit eine niedrigere Teststärke, wenn eine „gleichgerichtete“ Veränderung vorliegt. Gleichgerichtet heißt hier: Personen mit hoher Fähigkeit verbessern sich, und Personen mit niedriger

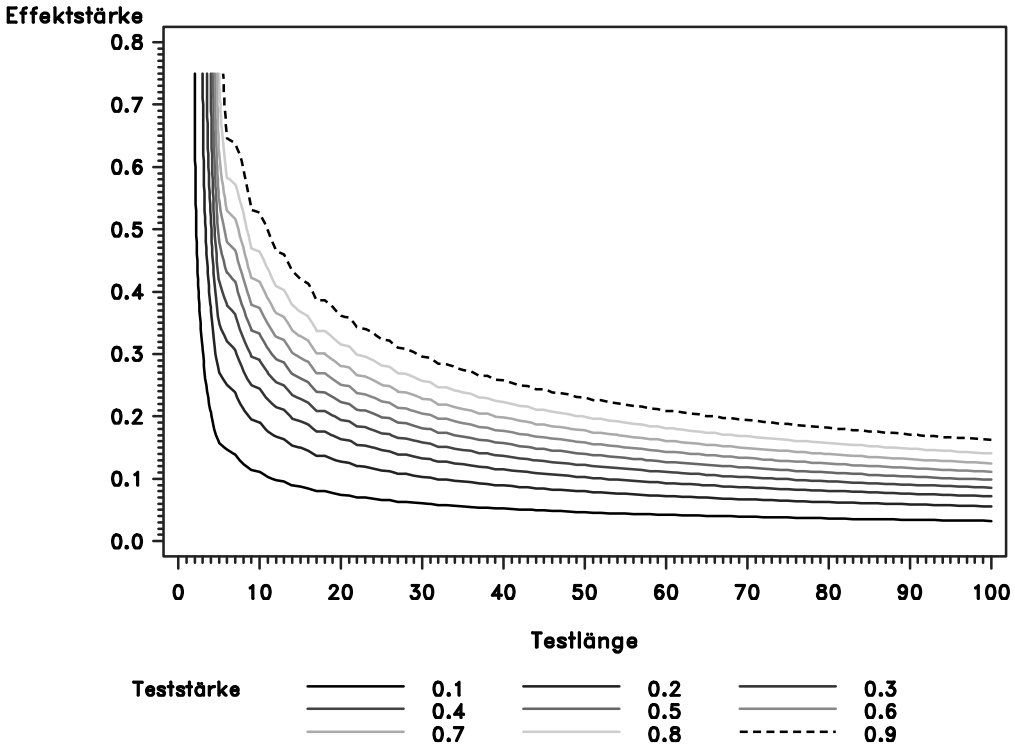

Abbildung 10.1: Teststärke in Abhängigkeit von Effektstärke und Zahl der Items bei $\alpha = 0.05$.

Fähigkeit verschlechtern sich. Eine niedrige Teststärke liegt weiterhin bei geringen Unterschieden zwischen Null- und Alternativhypothese vor.

Abbildung 10.2 zeigt die Teststärken bei einer Wahrscheinlichkeit für den Fehler erster Art von $\alpha = 0.01$, Abbildung 10.3 die Teststärken für $\alpha = 0.1$.

Bemerkenswert ist hieran vor allem, dass zwischen $\alpha = 0.01$ und $\alpha = 0.05$ ein deutlicher Anstieg der Teststärke bei konstanter Zahl eingehender Items sowie gegebener Effektstärke zu vermerken ist. Zwischen $\alpha = 0.05$ und $\alpha = 0.1$ besteht dagegen ein geringerer Unterschied hinsichtlich der Teststärke.

Abbbildung 10.4 zeigt den Zusammenhang zwischen dem Delta-Wert der Nullhypothese δ_0, dem tatsächlichen Delta-Wert δ_{akt} und der Effektstärke ES. Im Zusammenspiel mit den vorhergehenden Abbildungen kann man erkennen, dass die Teststärke erst bei relativ großen Abweichungen zufriedenstellend wird. So erreicht man z.B. eine Teststärke von 0.9 für 30 eingehende Items erst bei einem Unterschied von $\delta_{akt} - \delta_0 = 2$.

Die hier gezeigten Abbildungen helfen auch bei der Einordnung der Ergebnisse aus Kapitel 9. Dort wurde die Nullhypothese $\delta_{gt} = 0.2$ gegen die Alternative $\delta_{gt} \neq 0$ getestet. Es wurde zwischen 5 verzerrten und einer unverzerrten Gruppe von Probanden unterschieden. Innerhalb dieser Gruppen wurden weiterhin eine Kontrollgruppe, in der der tatsächliche Wert δ_{akt} des Veränderungsparameters um 0.2 schwankt, und eine Versuchsgruppe, in der

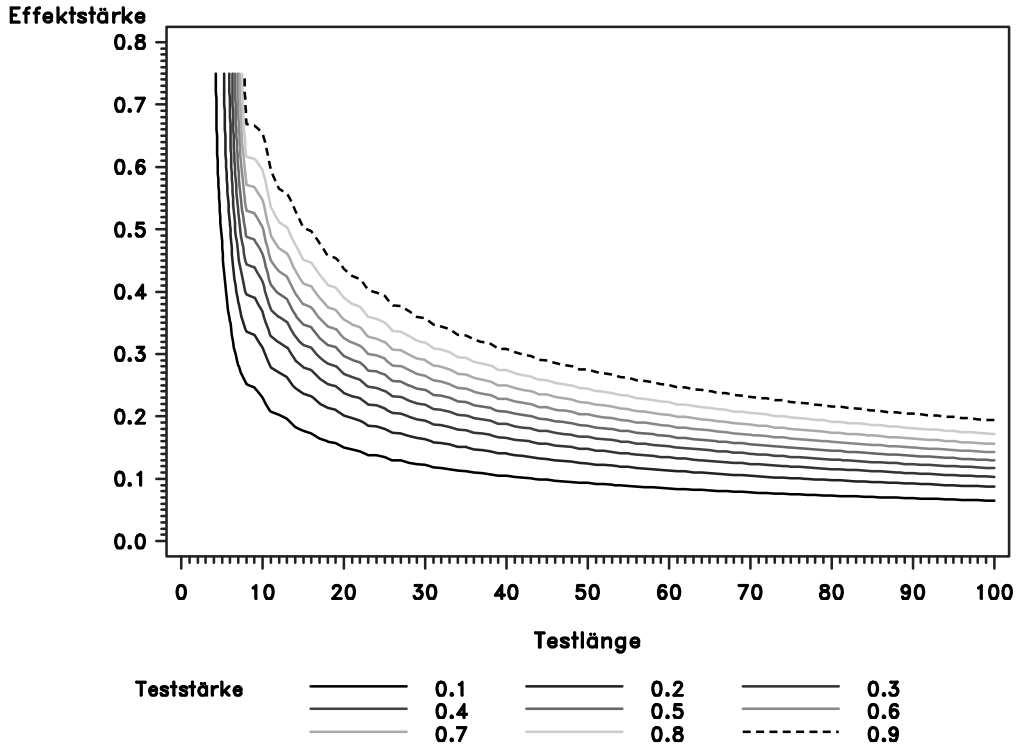

Abbildung 10.2: Teststärke in Abhängigkeit von Effektstärke und Zahl der Items bei $\alpha = 0.01$.

der tatsächliche Wert der Veränderung um -2 schwankt unterschieden. In den (verzerrten) Kontrollgruppen wurden i.d.R. wenige Ablehnungen verzeichnet.

Nach Betrachtung der Teststärken kann dieses Ergebnis folgendermaßen erklärt werden: Die Verzerrung nahm in den Gruppen 2 und 3 Maximalwerte von ± 2 an, meist (bei 60% der Probanden) vom Betrag her kleinere Werte als 1.6. In Gruppe 2 wurde die Wirkung der stärksten Fähigkeitsänderungen dadurch abgemildert, dass diese Maximalwerte für Personen mit extremen Fähigkeitswerten erreicht wurden. Dieser Effekt wird weiter unten im Text anhand der unbedingten Teststärkeberechnungen dargestellt. In Gruppe 4 betrug die maximale Fähigkeitsänderung durch die Verzerrung ca. 0.6. Sogar noch schwächer ausgeprägt war die Verzerrung in der Gruppe 5, in der die Schwierigkeit von 5 Items um 1.5 verändert wurde, sowie bei weiteren 5 Items um -1.5, so dass im Mittel keine Verzerrung bewirkt wurde.

Wenn man die Stärke dieser Verzerrungen mit den oben dargestellten Teststärken vergleicht, fällt auf, dass bei niedriger Zahl eingehender Items ($m \leq 20$) zumeist niedrige Teststärken vorliegen: Aus Abbildung 10.4 kann man ablesen, dass die Maximaleffekte in Gruppe 4 eine Effektstärke von höchstens 0.1 verursachen, was bei 20 Items mit einer Teststärke 0.15 gleichzusetzen ist, bei 30 Items mit einer Teststärke zwischen 0.15 und 0.30. Zumindest für die Gruppe 4 war die gewählte Verzerrung also zu klein, um regelmäßige Ablehnungen der Nullhypothese zu bewirken. Ein ähnliches Erklärungsmuster kann für die Ergebnisse in Gruppe 5 verwendet werden, wo im Mittel keine Verzerrung

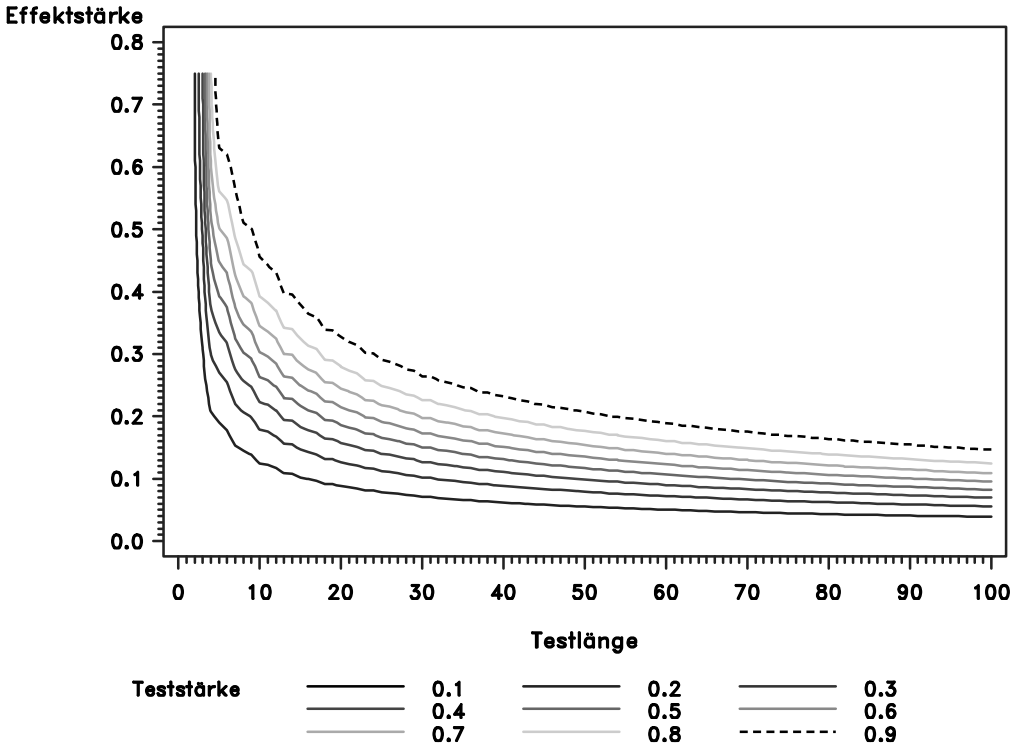

Abbildung 10.3: Teststärke in Abhängigkeit von Effektstärke und Zahl der Items bei $\alpha = 0.1$.

vorlag. Dagegen versagt dieses Erklärungsmuster bei den Gruppen 2 und 3, wo (für die maximalen Verzerrungen) Effektstärken von ca. 0.4 vorlagen, welche bei 30 Items zu einer Teststärke ≥ 0.9 und bei 20 Items zu einer Teststärke zwischen 0.7 und 0.8 führen sollten.

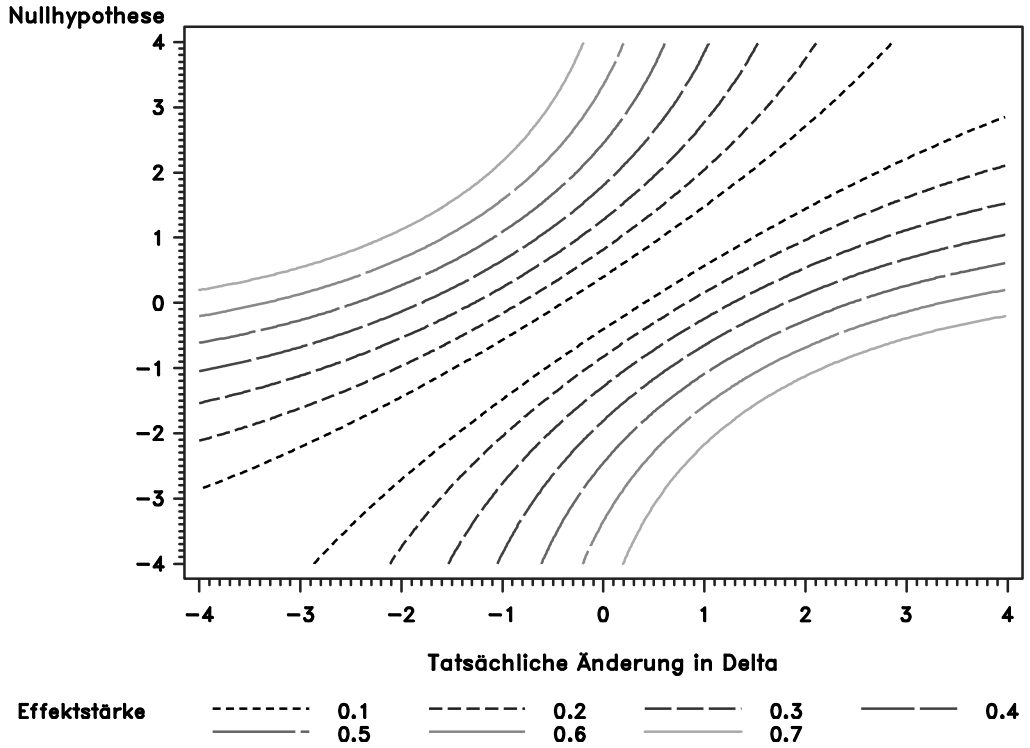

Abbildung 10.4: Effektstärke in Abhängigkeit vom Parameterwert der Nullhypothese und der Differenz zwischen Nullhypothese und tatsächlicher Veränderung.

10.2 Teststärke in Abhängigkeit vom Ausgangswert

10.2.1 Allgemeine Betrachtungen

Hier interessieren wir uns vor allem für das Verhalten der optimalen Personenfittests bei extremen Fähigkeitswerten im ersten Messzeitpunkt. Um hier Teststärken berechnen zu können, müssen wir die Zahl der verwendeten Items m als Realisation einer Zufallsgröße M auffassen. In Worten beschrieben ist M die Zahl der Items, bei denen eine vorgegebene Person v unterschiedliche Lösungen an den beiden Testzeitpunkten t_1 und t_2 erzielt. Für die Teststärke $P(X_{v\cdot 2} \in K|ES)$ können wir dann das folgende Modell annehmen:

$$P(X_{v\cdot 2} \in K|ES, \theta, \beta) = \sum_{m=0}^{N} P(M = m|ES, \theta, \beta) \times P(X_{v\cdot 2} \in K|m, ES) \quad . \tag{10.5}$$

N ist dabei die Zahl der einer Person an beiden Zeitpunkten vorgelegten Items.

$P(X_{v\cdot 2} \in K|m, ES)$ ist die gleiche Größe, deren Plots im vorangegangenen Abschnitt dargestellt wurden. Eine Berechnung dieser Größe ist daher unproblematisch. Für die Berechnung von $P(M = m|ES, \theta, \beta)$ hingegen müssen wir die Schwierigkeitsparameter β der Items kennen. Dann erhalten wir

$$P(M = m|\delta_{akt}, \delta_0, \theta_v) = \sum \prod_i P[(X_{vi1} = 1 \cup X_{vi2} = 0) \cap (X_{vi1} = 0 \cup X_{vi2} = 1)] \quad , \tag{10.6}$$

wobei über alle Teilmengen mit genau m richtig beantworteten Items summiert wird.

10.2.2 Untersuchung des idealtypischen Verhaltens der Teststärke bei extremen Fähigkeiten

Als Beispiel für das Verhalten der Teststärke bei extremen Fähigkeitswerten wollen wir den Fall sehr hoher Fähigkeiten θ_v untersuchen. Dabei sind zwei Typen von Items zu unterscheiden: Extrem schwierige Items, bei denen $\beta_{it} \approx \theta_v$ gilt, und normal schwierige Items mit $\beta_{it} \ll \theta_v$. Bei letzteren Items ist $P[(X_{vi1} = 1 \cup X_{vi2} = 0) \cap (X_{vi1} = 0 \cup X_{vi2} = 1)] \approx 0$. Nun sind zwei Fälle zu unterscheiden:

- Die Fähigkeit wird durch die Veränderung erhöht: In diesem Fall ist es ausreichend, sich auf die Gruppe extrem schwieriger Items zu beschränken, um eine idealtypische Verteilung zu konstruieren, da bei allen Itemteilmengen mit „normal schwierigen“ Items
$$\prod_i P[(X_{vi1} = 1 \cup X_{vi2} = 0) \cap (X_{vi1} = 0 \cup X_{vi2} = 1)] \approx 0 \tag{10.7}$$
gilt. Die Kombination extrem hoher Fähigkeiten und einer verstärkenden Veränderung erniedrigt daher die Zahl der in den Personenfittest eingehenden Items.

- Die Fähigkeit wird durch die tatsächliche Veränderung erniedrigt. In diesem Fall besteht die Menge der Items, bei denen eine Lösung in nur einem Zeitpunkt überhaupt möglich ist, aus den extrem schwierigen Items und aus einer Teilmenge der normal schwierigen Items, die auch jetzt mit einer deutlich von 0 verschiedenen Wahrscheinlichkeit nicht gelöst werden können. Der Effekt extrem hoher Fähigkeiten ist bei einer abschwächenden Veränderung daher kleiner als bei einer verstärkenden Veränderung.

10.2.3 Berechnung der Teststärke bei extremen Fähigkeitswerten für die durch die Simulation aus Kapitel 9 vorgegebene Itemstruktur

Teststärke bei fehlendem funktionalem Zusammenhang zwischen Fähigkeit und tatsächlicher Veränderung

Zunächst werden wir die Abhängigkeit zwischen Fähigkeit und Teststärke darstellen, wenn wir keinen funktionalen Zusammenhang zwischen der tatsächlichen Veränderung δ_{akt} und der Fähigkeit annehmen. Die nachfolgende Grafik 10.5 zeigt die Teststärke bei gegebener Fähigkeit in Abhängigkeit von der tatsächlichen Veränderung δ_{akt}.

Die Teststärke wird gemäß den Formeln (10.5) und (10.6) berechnet. Ein einfacher Algorithmus für (10.6) ergibt sich folgendermaßen: Mit Hilfe der Substitution

$$P[(X_{vi1} = 1 \cup X_{vi2} = 0) \cap (X_{vi1} = 0 \cup X_{vi2} = 1)] = \frac{\exp(\omega_i)}{exp(1 + \omega_i)} \tag{10.8}$$

sowie mit

$$\begin{aligned} z_i &= 1 - [x_{vi1}x_{vi2} + (1 - x_{vi1})(1 - x_{vi2})] \\ P(Z_i = 1) &= P[(X_{vi1} = 1 \cup X_{vi2} = 0) \cap (X_{vi1} = 0 \cup X_{vi2} = 1)] \end{aligned} \tag{10.9}$$

erhält man

$$\begin{aligned} P(M = m|\delta_{akt}, \delta_0, \theta_v) &= \sum \prod_i \frac{\exp(z_i\omega_i)}{exp(1 + \omega_i)} \\ &= \frac{\sum \prod_i \exp(z_i\omega_i)}{\prod_i (exp(1 + \omega_i))} \\ &= \frac{\gamma_m^N(\omega_1, \ldots, \omega_N)}{\prod_i (1 + \omega_i)} \quad . \end{aligned} \tag{10.10}$$

Hierbei ist $\gamma_m^N(\omega_1, \ldots, \omega_N)$ die elementare symmetrische Funktion der $\exp(\omega_i)$ mit m richtig beantworteten Items und N Items insgesamt. $\gamma_m^N(\omega_1, \ldots, \omega_N)$ kann leicht mittels des sog. Summationsalgorithmus berechnet werden (vgl. z.B. [Fischer, 1995a]).

Als Beispiel für die Wirkung der Fähigkeit im ersten Messzeitpunkt auf die Teststärke wollen wir die in der Simulation des Kapitels 9 definierte Itemstruktur verwenden. Diese bestand in beiden Messzeitpunkten aus 30 Items mit Schwierigkeiten -1.4, -1.3,...,0, 0.1, ..., 1.5. Als Nullhypothese wurde $\delta_0 = 0.2$ verwendet.

Abbildung 10.5 verdeutlicht den Zusammenhang zwischen Teststärke und Fähigkeit bei der gewählten Itemstruktur. Bei extremen Fähigkeitswerten mit $|\theta| >= 3.1$ ist die Teststärke stets kleiner als 0.1, wenn Veränderung und Fähigkeit das gleiche Vorzeichen aufweisen. Bei ungleichem Vorzeichen ist dagegen eine wesentlich höhere Teststärke zu beobachten. Die größten Teststärken sind rund um $\theta = 0$ zu beobachten. Dabei liegt das eigentliche Maximum der Teststärke (bei vorgegebener Veränderung) stets so, dass das Vorzeichen der Fähigkeit ungleich dem Vorzeichen der Veränderung ist.

Als Bereich hoher Teststärken definieren wir jetzt (willkürlich) alle Teststärken, die größer als 0.6 sind. Große Teststärken treten vor allem dann auf, wenn

- die Fähigkeit zwischen -1.5 und 3 liegt und die Veränderung kleiner als -2 ist, oder
- die Fähigkeit zwischen -3 und 1.5 liegt und die Veränderung größer ist als 2.

Die Abbildung 10.5 liefert eine Erklärung für das Zustandekommen der Ablehnhäufigkeiten für die Abweichungsgruppen 2 und 3 aus der Simulation in Kapitel 9. In Abweichungsgruppe 2 hängt die tatsächliche Veränderung positiv linear von der Fähigkeit der Versuchspersonen ab. Dies führt bei Versuchspersonen mit extremen Fähigkeiten zu niedrigen Teststärken, woraus wiederum eine niedrige relative Ablehnhäufigkeit in der Simulation aus Kapitel 9 folgt.

In Abweichungsgruppe 3 wird ein negativer linearer Zusammenhang zwischen Fähigkeit und Veränderung vorausgesetzt. Dieser führt zu einer relativen Ablehnhäufigkeit, die meist höher als in allen anderen Kontrollgruppen ist, jedoch niedriger als in allen Versuchsgruppen. Dies ist durch zwei Effekte zu erklären:

- Rund um $\theta = 0$ ist die Veränderung am schwächsten. Dies ist jedoch gleichzeitig ein Bereich mit hohen Teststärken. Somit fehlt ein möglicher Fähigkeitsbereich im Vergleich zu einer konstanten Veränderung, wie sie in allen Versuchsgruppen vorliegt.
- Wie man aus Abbildung 10.5 ablesen kann, sinkt die Teststärke, wenn man bei negativer Veränderung eine höhere Fähigkeit als 1.5 besitzt, bzw. wenn man bei positiver Veränderung eine niedrigere Fähigkeit als -1.5 aufweist. Dies sind aber die Bereiche mit der vom Betrag her größten Veränderung in der Abweichungsgruppe 3. Die stärksten Veränderungen liegen hier also nicht einem Bereich mit der höchsten unbedingten Teststärke, was dazu führt, dass vorhandene Abweichungen zu selten erkannt werden.

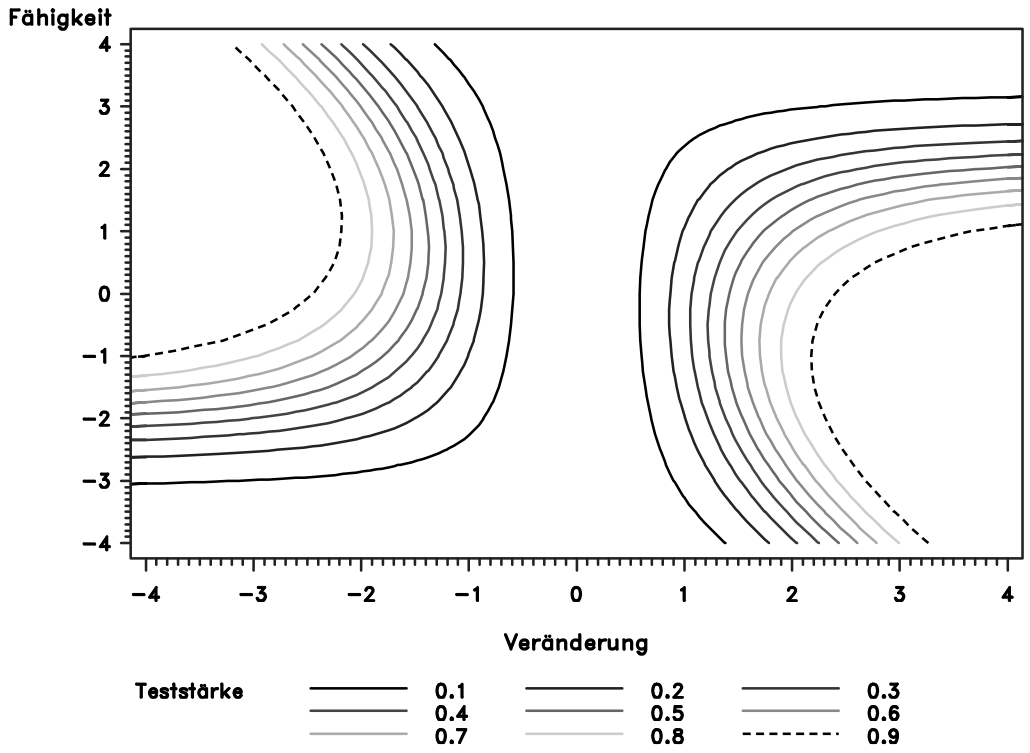

Abbildung 10.5: Teststärke in Abhängigkeit von Fähigkeit und tatsächlicher Veränderung bei $\alpha = 0.05$. Angenommene Itemstruktur wie im Text beschrieben.

In den Versuchsgruppen hingegen war stets eine Abweichung von -2.2 gegenüber der Nullhypothese vorausgesetzt. Eine solch große Abweichung führt gemäß Abbildung 10.5 in den Fähigkeitsbereichen $[-1.5; 3]$ zu Teststärken größer als 0.6. Dies erklärt das gute Abschneiden der Tests in allen Versuchsgruppen.

10.3 Verhalten des Binomialtests bei mehrdimensionalen Alternativen

In diesem Abschnitt wird die Teststärke bei mehrdimensionalen Alternativen untersucht. Als „mehrdimensionale Alternative“ verstehen wir dabei die folgende Situation: Getestet wird

H_0: Für alle Items existiert nur ein Veränderungsparameter δ_{g2}, und es gilt: $\delta_{g2} = c$,

gegen:

H_A: Es gibt im zweiten Untersuchungszeitpunkt mehrere Veränderungsparameter $\delta_{12}^{(1)}, \ldots, \delta_{g2}^{(l)}$, die jeweils nur für eine Teilmenge der Items zutreffen. Für mindestens einen Veränderungsparameter $\delta_2^{(h)}$ gilt: $\delta_{g2}^{(h)} \neq c$.

Dies ist die gleiche Testsituation wie in Kapitel 6.3 für die Untersuchung der Nullhypothese

$$\delta_{12}^{(1)} = \ldots = \delta_{g2}^{(l)} = c \ .$$

Wie in 6.3 nehmen wir an, dass für jeden Veränderungsparameter $\delta_{g2}^{(h)}$ eine Menge $I_h = \left\{i_1^h, \ldots, i_{l_h}^h\right\}$ von Items $i_1^h, \ldots, i_{l_h}^h$ existiert, für die $\delta_{g2}^{(h)}$ der wahre Veränderungsparameter ist.

Wenn die obige Alternative zutrifft, besitzt die Testgröße $X_{v\cdot 2}$ die bedingte Verteilung

$$\begin{aligned} Pr\left(X_{v\cdot 2} = x_{v\cdot 2} \mid x_{v1\cdot}, \ldots, x_{vm\cdot}\right) = \\ = \frac{\sum \exp\left[\sum_k \sum_{i \in I_k} \left(x_{vi2}\delta_{k2}\right)\right]}{\sum \exp\left[\sum_k \sum_{i \in I_k} \left(x_{vi2}\delta_{k2}\right)\right]} \ . \end{aligned} \tag{10.11}$$

Im Nenner dieses Bruchs wird über alle Antwortmuster mit vorgegebenen suffizienten Statistiken $x_{v1\cdot}, \ldots, x_{vm\cdot}$ summiert. Im Zähler des Bruchs wird über alle Antwortmuster mit $X_{v\cdot 2} = x_{v\cdot 2}$ und vorgegebenen suffizienten Statistiken summiert.

Die Berechnung dieser Wahrscheinlichkeit kann auf folgende Weise vereinfacht werden:

- Man berücksichtigt nur solche Items, die nur an einem Zeitpunkt richtig beantwortet wurden. Terme von Items, die an beiden Zeitpunkten gleich beantwortet wurden, kürzen sich heraus.

- Nach dem Kürzen solcher Terme besitzt der Zähler aus (10.11) die Struktur einer elementaren symmetrischen Funktion in dem n_1-dimensionalen Vektor

$$\Delta = \left(\delta_{g2}^{h_1}, \ldots, \delta_{g2}^{h_k}\right)$$

der zu den Items $i_1, \ldots, i_{n_1}$ gehörenden Veränderungsparameter. Dabei sind $i_1, \ldots, i_{n_1}$ die Items, die nur an einem Messzeitpunkt richtig gelöst wurden.

- Der Nenner aus (10.11) ist dann die Summe aller solcher elementaren symmetrischen Funktionen.

Mit diesen Hilfsmitteln berechnen wir die (bedingte) Teststärke unseres Binomialtests. Dabei legen wir folgende Voraussetzungen fest:

- Wir gehen von $m = n_1 = 30$ eingehenden Items aus.
- Als Veränderung unter der Nullhypothese legen wir stets $\delta_{g2} = 0$ fest.
- Die Zahl der unter der Alternativhypothese auftretenden Veränderungsparameter wird variiert. Es werden zwei unterschiedliche Situationen untersucht: Zum einen ein Design mit 3 verschiedenen Veränderungsparametern, wobei eine Gruppe immer der Nullhypothese entspricht. Die Teststärkeberechnungen werden dabei für unterschiedliche Gruppengrößen durchgeführt.
  Bei einem zweiten Design wird 1 Veränderungsparameter pro Item angenommen. Der Einfachheit halber wird bei diesem Design weiterhin angenommen, dass die Werte der Veränderungsparameter gleichmäßig über ein vorgegebenes Intervall verteilt sind.
- Das arithmetische Mittel der in Δ enthaltenen Veränderungsparameter wird variiert. Bei Untersuchungsdesign 1 (= 3 Gruppen) variiert der Wert des Veränderungsparameters in den von der Nullhypothese abweichenden Gruppen zwischen -2 und 2. Das arithmetische Mittel der nicht der Nullhypothese entsprechenden Veränderungswerte liegt daher ebenfalls in diesem Bereich.
  Für Design 2 wird das Intervall variiert, in dem die Untersuchungsparameter liegen. Als Minimum des Intervalls werden Werte zwischen -2 und 2 vorgegeben, die Intervalllänge variiert zwischen 0 und 2.
- Bei Design 1 wird die Zahl der nicht der Nullhypothese entsprechenden Items variiert. Dabei werden beide Abweichungsgruppen mit gleicher Größe gewählt. Untersucht wird die Teststärke bei 3 Items pro Abweichungsgruppe und bei 9 Items pro Abweichungsgruppe.
- Als Signifikanzniveau wird stets $\alpha = 0.05$ verwendet.

Die folgenden Diagramme stellen die berechneten Teststärken dar. Die Abbildungen 10.6 und 10.7 zeigen das Verhalten der Teststärke bei einer der Nullhypothese entsprechenden und zwei von der Nullhypothese abweichenden Gruppen von Items. Die X-Achse zeigt die Veränderung in einer der beiden Abweichungsgruppen, die Y-Achse die Veränderung

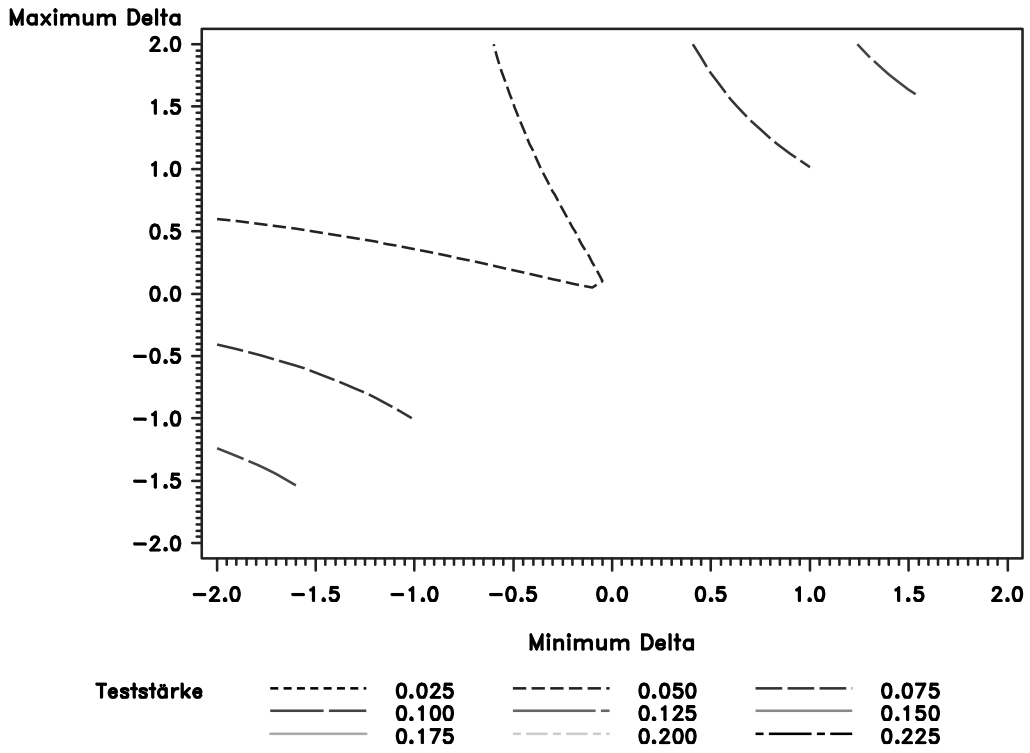

Abbildung 10.6: Teststärke bei mehrdimensionaler Veränderung und $\alpha = 0.05$. Drei Veränderungsparameter. Insgesamt 6 Items unterscheiden sich von der Nullhypothese.

in der anderen Gruppe. Dargestellt wird nur der Bereich, für den die Veränderung der Y-Gruppe größer als die Veränderung der X-Gruppe ist.

Am auffälligsten in Abb. 10.6 ist die stets sehr niedrige Teststärke. In dieser Abbildung weisen nur 6 von 30 Items eine im Vergleich zur Nullhypothese unterschiedliche Veränderung auf. Weiterhin ist bemerkenswert, dass die Teststärke nahezu Null wird, wenn das arithmetische Mittel der Veränderungen ungefähr beim Wert 0 – und somit in Nähe der Nullhypothese – liegt.

Bei Abb. 10.7 weisen 18 von 30 Items eine andere Veränderung auf als die in der Nullhypothese vorgegebene. Dies macht sich in einer deutlich erhöhten Teststärke bemerkbar. In dem untersuchten Bereich liegt das Maximum der Teststärke bei 0.7. Allerdings findet man rund um die Gerade $Y = -X$ einen Bereich mit äußerst niedriger Teststärke. In diesem Bereich steht man vor der folgenden Situation: Es existiert eine Gruppe von Items, bei der die Veränderung der Nullhypothese entspricht, sowie zwei Gruppen von Items, deren Veränderungen $\delta^1_{g2}, \delta^2_{g2}$ sich von der Nullhypothese unterscheiden. In dem angesprochenen Bereich rund um die Gerade $Y = -X$ gilt $\delta^1_{g2} \approx (-1) \times \delta^2_{g2}$, so dass die von diesen Veränderungen beeinflussten Antworten sich gegenseitig auslöschen. Einen solchen Zustand werden wir im Folgenden auch als „um die Nullhypothese symmetrische Abweichungen“ bezeichnen.

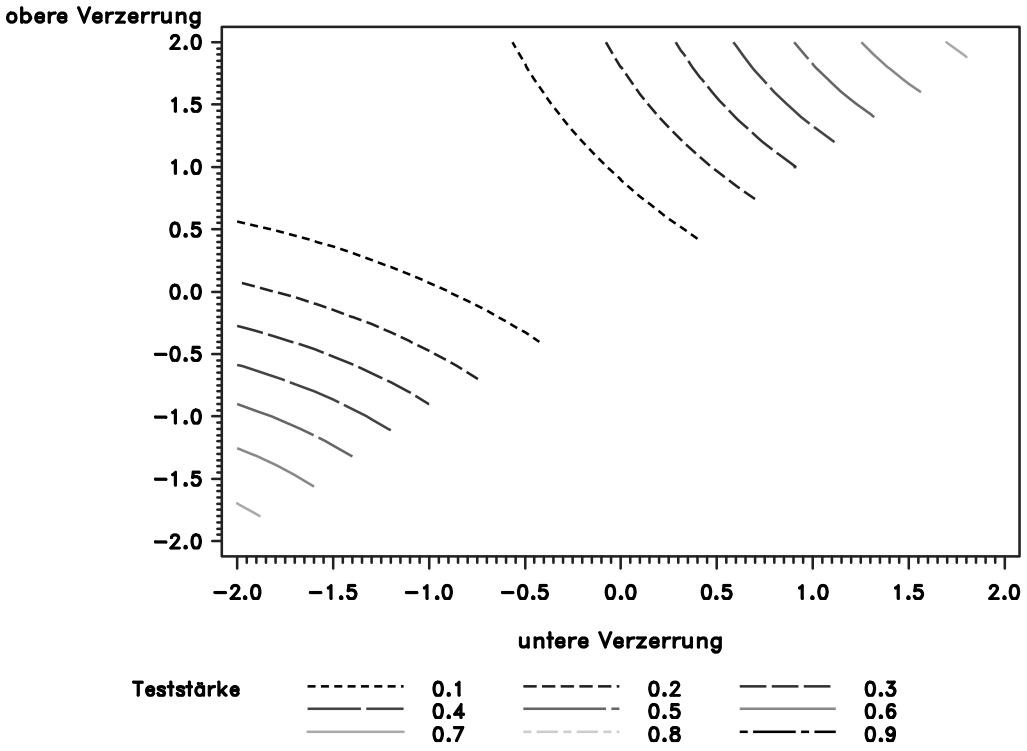

Abbildung 10.7: Teststärke bei mehrdimensionaler Veränderung und $\alpha = 0.05$. Drei Veränderungsparameter insgesamt. 18 Items unterscheiden sich von der Nullhypothese.

In Abbildung 10.8 ist eine andere Form der Abweichung von der Nullhypothese zugrunde gelegt: Hier existiert für jedes Item eine eigene Veränderung, und die Veränderungen sind gleichmäßig über ein Intervall $[\Delta_{min}; \Delta_{min} + l]$ der Länge l verteilt. Somit wird hier die Teststärke in Abhängigkeit von kleinstem Veränderungswert und Länge des kleinsten, die Veränderungen überdeckenden Intervalls dargestellt. Die Intervalllänge ist dabei so gewählt, dass die maximale Intervalllänge gleich dem größtmöglichen Unterschied zwischen den Veränderungen der Abweichungsgruppen ist.

Der Unterschied zwischen den Abbildungen 10.6 und 10.7 einerseits, und der Abbildung 10.8 andererseits besteht also darin, dass bei den Abbildungen 10.6 und 10.7 die abweichenden Veränderungen auf nur zwei Punkte $\left\{\delta^1_{g2}, -\delta^1_{g2}\right\}$ konzentriert sind, während die abweichenden Veränderungen bei 10.8 eher gleichmäßig verteilt sind. Auffällig beim Vergleich der drei Abbildungen ist, dass bei 10.8 wesentlich höhere Teststärken auftreten. Zudem ist der Bereich sehr niedriger Teststärke recht schmal. Der Bereich niedriger Teststärke beschränkt sich auf Fälle, bei denen die Intervalle symmetrisch um die Nullhypothese liegen.

Zudem ist eine gewisse Vergrößerung des Bereichs niedriger Teststärke bei großen Intervalllängen zu erkennen. Eine große Schwankungsbreite der Veränderungen wirkt sich also negativ auf die Teststärke aus. Sonst ist bemerkenswert, dass die „Höhenlinien“ gleicher Teststärke nahezu parallel liegen. Zumindest im dargestellten Bereich steigt die Teststärke nahezu linear bezüglich der Länge des Wertebereichs der Veränderung/der maximalen

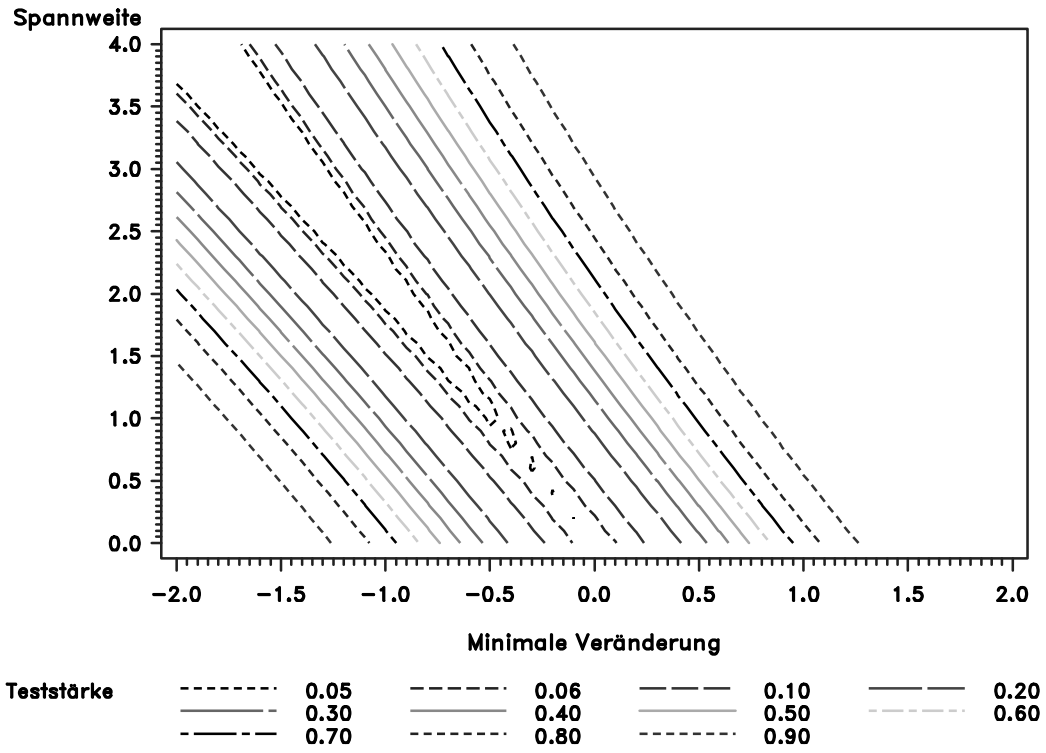

Abbildung 10.8: Teststärke bei mehrdimensionaler Veränderung und $\alpha = 0.05$. Ein Veränderungsparameter pro Item. Y-Achse: Länge des Intervalls, in dem die Veränderungswerte liegen. X-Achse: Minimum des Intervalls.

Veränderungsdifferenz an.

Zusammenfassung

Die hier gezeigten Diagramme lassen Rückschlüsse auf das Verhalten der einfachen Binomialtests bezüglich multinomialer Veränderung zu. Deutlich wird, dass symmetrische Abweichungen nur schlecht zu erkennen sind. Verstärkt wird dieser Effekt, wenn die Veränderungen der Alternativhypothese auf nur wenige Punkte konzentriert sind.

Eine weite Streuung führt bei gleich bleibendem Maximalabstand zur Nullhypothese zu geringeren Teststärken. Dies ist als deutlicher Mangel der hier vorgestellten Tests zu werten. Statt der Binomialtests sollte in diesem Fall entweder einer der in Kapitel 6.3 vorgestellten multivariaten Tests verwendet werden, oder man sollte sich auf eine Gruppe von Items mit vermutetem Item Bias beschränken, und für diese Gruppe von Items den Binomialtest durchführen. Letztere beide Verfahren können als Alternativverfahren zu den in Kapitel 3 vorgestellten Methoden angesehen werden.

Wie in der Teststärkenuntersuchung gezeigt, ist die Verwendung von Personenfittests wenig erfolgversprechend, wenn bei den Bias-behafteten Items mehrere Veränderungswirkun-

gen auftreten, deren arithmetisches Mittel ungefähr gleich dem Veränderungswert der Nullhypothese ist. Dieser Effekt ist bei den Tests aus Kapitel 3 eher unwahrscheinlich: Bei den dortigen Tests können aufgrund der Konstruktion des Testverfahrens entgegengerichtete Veränderungen bei einzelnen Items einander nicht auslöschen.

Gemäß den Ergebnissen der Teststärkeuntersuchungen existieren weitere Kritikpunkte:

- Die Richtung der Veränderung im Zusammenspiel mit der Fähigkeit beeinflusst die Güte der Personenfittests. Z.B. resultiert bei Personen mit hohen Fähigkeiten eine niedrigere Teststärke, wenn die Verbesserung auftritt, als bei Personen mit niedriger Fähigkeit. Dieser Kritikpunkt trifft aber auch auf die in Kapitel 3 vorgestellten Tests zu: Alle Testverfahren der Veränderungsmessung weisen eine Ausgangswertabhängigkeit auf (vgl. z.B. [Krause, 1997]).
- Um mit Personenfittests eine hohe Teststärke erzielen zu können, benötigt man eine große Zahl von Items, die in den beiden Messzeitpunkten unterschiedlich beantwortet wurden. Daher sind Personenfittests für Veränderung auf Untersuchungen mit einer großen Zahl von Items spezialisiert. Genau das Gegenteil ist bei den Tests aus Kapitel 3 der Fall. Diese Tests sind nur dann sinnvoll einzusetzen, wenn eine niedrige Itemzahl vorliegt: Diese Untersuchungsmethoden modellieren die bedingte Wahrscheinlichkeit $Pr(X_{gl.}\,|R_{t_1} = r_1, R_{t_2} = r_2\,)$, die für alle aus den Kombinationen der Summenscores R_{t_1} und R_{t_2} gebildeten Zellen berechnet wird. Eine hohe Itemanzahl führt zu einer hohen Zahl von Zellen, und dies wiederum zu einer hohen Anzahl benötigter Versuchspersonen. Bei einer hohen Itemzahl ist von den Methoden aus Kapitel 3 eher abzuraten.

Kapitel 11

Untersuchungen zur Konstanz der Veränderung beim Lebensqualitätsfragebogen EORTC QLQ-C30

Im folgenden Kapitel führen wir anhand eines realen Datensatzes vor, welche Möglichkeiten sich durch die Analyse mittels optimalen Personenfittests für die Veränderungsmessung ergeben.

11.1 Aufbau des EORTC QLQ-C30

Der Datensatz dieses Kapitels stammt aus einer Untersuchung zur Veränderung der Lebensqualität an 296 Krebspatienten, die an der Charité im Jahr 2001 von Dipl.-Psych. Stefan Künstner im Rahmen seiner Dissertation durchgeführt (vgl. [Künstner, 2002]) wurde. Der aus dieser Untersuchung hervorgegangene Datensatz enthält die Antworten auf die Items des EORTC QLQ-C30 bei 2 Messzeitpunkten (vgl. Fragebogen im Anhang), sowie einige weitere Variablen wie Alter, Geschlecht und Zeitraum zwischen den Messzeitpunkten. Als Fragebogen zur Lebensqualität wird der EORTC QLQ-C30 in der Version 3.0 verwendet. Dieser Fragebogen besteht aus 28 vierstufigen Items sowie 2 siebenstufigen Items.

[Aaronson et al., 1993] beschreibt die psychometrischen Eigenschaften des EORTC QLQ-C30: Der EORTC QLQ-C30 ist ein multidimensionaler Fragebogen, der sich aus 6 Teilskalen zusammensetzt, die die Leistungsfähigkeit abfragen, sowie 3 Teilskalen und 6 Items, die einzelne Symptome abfragen. Als Subskalen verwenden lassen sich:

- Items 1-5: „Physische Beeinträchtigungen“ (bzw. Physis)
- Items 6, 7: „Rollenverhalten“
- Items 20, 25: „Kognitive Auswirkungen“
- Items 21-24: „Emotionale Auswirkungen“
- Items 26, 27: „Soziale Beeinträchtigungen“
- Items 29, 30: „Globale Lebensqualität“
- Items 10, 12, 18: „Erschöpfung“
- Items 14, 15: „Erbrechen und Übelkeit“
- Items 9, 19: „Schmerzen“

Die übrigen Items des EORTC QLQ-C30 sind symptombezogen (z.B. Frage 8: „Waren Sie kurzatmig“) und lassen sich keiner eigenen Subskala zuordnen.

[Aaronson et al., 1993] sehen alle Teilskalen bis auf die Teilskala „Rollenverhalten“ als hinreichend reliabel an (Cronbachs $\alpha \geq 0.70$ bei mindestens einem Messzeitpunkt). Dazu bleibt anzumerken, dass dies Minimalanforderungen an die Reliabilität sind. Kritisch ist nach Meinung des Autors, dass es ausreicht, wenn die Bedingung für Cronbachs α an einem Messzeitpunkt erfüllt wird. Dieses Vorgehen führt zu einem größeren „Fehler erster Art“ bei der Erkennung reliabler Skalen: Die Wahrscheinlichkeit, dass eine Skala als reliabel beurteilt wird, obwohl sie dies nicht ist, kann sich durch das Vorgehen von [Aaronson et al., 1993] beträchtlich erhöhen. Falls man nur solche Skalen als reliabel ansieht, bei denen an beiden Messzeitpunkten Cronbachs $\alpha \geq 0.70$ gilt, so sind nur noch die emotionale Teilskala, die globale Lebensqualität, die Erschöpfungs- und die Schmerzskala reliabel.

Wegen der ausgeprägten Multidimensionalität des Fragebogens und der recht geringen Inter-Skalenkorrelationen ist von der Verwendung eines Gesamtsummenscores für diesen Fragebogen wohl abzuraten (vgl. auch [Aaronson et al., 1993], S. 369):

> In general, the inter-scale correlations were only of a moderate size indicating that, although related, they are assessing distinct components of the quality-of-life construct.

Der zu Verfügung stehende Datensatz wird auf zwei Fragestellungen hin untersucht: In Abschnitt 11.2 zeigen wir, wie man mit den in dieser Arbeit beschriebenen Methoden einzelne Teilskalen des Tests zusammenfassen und so zu einer Gesamtaussage bezüglich der Veränderung kommen kann. Außerdem zeigen wir, wie man mit den hier vorgestellten Methoden untersuchen kann, ob für alle Schwellen eines Items die gleiche Veränderungswirkung c angenommen werden kann (vgl. auch Kapitel 11.3).

11.2 Zusammenfassende Aussagen über mehrere Teilskalen

Mehrdimensionale Personenfittests können für zusammenfassende Aussagen über die Veränderung bei mehreren Teilskalen verwendet werden. Dies werden wir am Beispiel des EORTC QLQ-C30 demonstrieren.

Theorie

Mit Hilfe von Personenfittests ist es möglich, zusammenfassende Aussagen über Veränderungswirkungen zu machen, nämlich durch Ablehnung einer Nullhypothese der Art

> H_0: In keinem der untersuchten Subtraits kann eine Veränderung festgestellt werden.

Effektgrößen erlauben standardisierte Aussagen über die Stärke der Veränderung bei einer Person bzw. einer Subpopulation.

Realisation

Anhand des EORTC QLQ-C30-Datensatzes wird untersucht, ob bei mindestens einer der Subskalen Physis, Rollenverhalten, Emotionale Auswirkungen und Erschöpfung eine Veränderung zwischen den beiden Messzeitpunkten festgestellt werden kann.

Dazu werden für jede Itemantwort die zugehörigen dichotomen Indikatorvariablen c_{vijt} der einzelnen Schwellen berechnet. Hierbei steht v für die Person, i für ein Item, j für die Schwelle eines Items und t für den Messzeitpunkt. Diese Indikatorvariablen werden dann als Grundlage für einen mehrdimensionalen Personenfittest bezüglich der Nullhypothese

$$H_0: \quad \delta_{12} = \delta_{22} = \delta_{32} = \delta_{42} = 0$$

verwendet. Dabei ist δ_{12} der Veränderungsparameter im 2. Messzeitpunkt des Subtraits Physis, δ_{22} der Veränderungsparameter im 2. Messzeitpunkt des Subtraits Rollenverhalten, δ_{32} der Veränderungsparameter im 2. Messzeitpunkt des Subtraits Emotionale Auswirkungen und δ_{42} der Veränderungsparameter im 2. Messzeitpunkt des Subtraits Erschöpfung.

Nicht alle der 296 Versuchspersonen, aus denen die Stichprobe besteht, können für die Personenfittests verwendet werden:

- Bei 84 Personen fehlen eine oder mehrere Antworten.
- Bei weiteren 144 Personen gehen weniger als 8 Antworten in eine 4×2-Kontingenztafel ein, d.h. die durchschnittliche Zellenbesetzung dieser Kontingenztafel ist kleiner als 1. Dies ist auch bei einem exakten Test eine für sinnvolle Aussagen zu niedrige Zellenbesetzung.

Somit werden die Personenfittests an 68 Personen durchgeführt. Bei 52 Personen kann die Nullhypothese abgelehnt werden. Der Adding-Of-Logs-Test auf Gesamtsignifikanz führt zu einer Testgröße von 3757.82 bei 424 Freiheitsgraden. Der P-Wert der zugehörigen χ^2-Verteilung ist kleiner als 1×10^{-12}. Somit kann zum Gesamsignifikanzniveau von 0.05 die Hypothese

> H_0^{Gesamt}: Alle Ablehnungen der Einzelnullhypothesen sind zufällig zustande gekommen

abgelehnt werden. Die Gesamtsignifikanzprüfung mit dem Binomialtest führt zu ähnlichen Ergebnissen. Der Wert 52 der Testgröße führt bei der $B(212; 0.1)$-Verteilung zur Ablehnung der Gesamtnullhypothese mit einem P-Wert, der kleiner als 1×10^{-12} ist.

In einem weiteren Schritt werden die Effektgrößen für die Veränderung der einzelnen Traits gemessen. Falls in einem Trait keine Schwelle auftritt, die genau einmal beantwortet wird, wird für die Effektgröße der Wert 0 angenommen. Da wir Beträge von Effektgrößen untersuchen, ist dies der kleinstmögliche Wert, den ein Effekt annehmen kann. Als Effektgröße verwenden wir die von [Cohen, 1969], Kap. 6.10 für den Biomialtest vorgeschlagene Effektgröße

$$x = \left| \arcsin\left(\sqrt{\hat{p}}\right) - \arcsin\left(\sqrt{p_0}\right) \right| \sqrt{m} \quad . \tag{11.1}$$

Hierbei ist $\hat{p}$ der Anteil der nur im zweiten Messzeitpunkt gelösten Items an den in genau einem Zeitpunkt gelösten Items, p_0 der durch die Nullhypothese vorgegebene Wert dieses Anteils, und m die Zahl der Items, die in genau einem Zeitpunkt gelöst wurden. Zunächst wird die Größenordnung der Effektstärken (falls diese berechenbar waren) für die einzelnen Subskalen mittels Boxplots dargestellt (vgl. 11.1). Die Boxplots zeigen die Effektstärken für alle Personen, bei denen ein Test durchgeführt wurde.

Wie man sieht, weisen die Subskalen Physis, Emotionale Auswirkungen und Erschöpfung ähnliche Verteilungen auf. Der Median der Verteilungen liegt stets bei ca. 2.2. Auch weisen die Verteilungen in diesen drei Fällen annähernd gleiche Streuung auf (gemessen durch den Interquartilsabstand). Abweichend davon ist lediglich der Boxplot für das Rollenverhalten. Dieser weist eine starke Konzentration zwischen den Werten 2.1 und 3.0 auf.

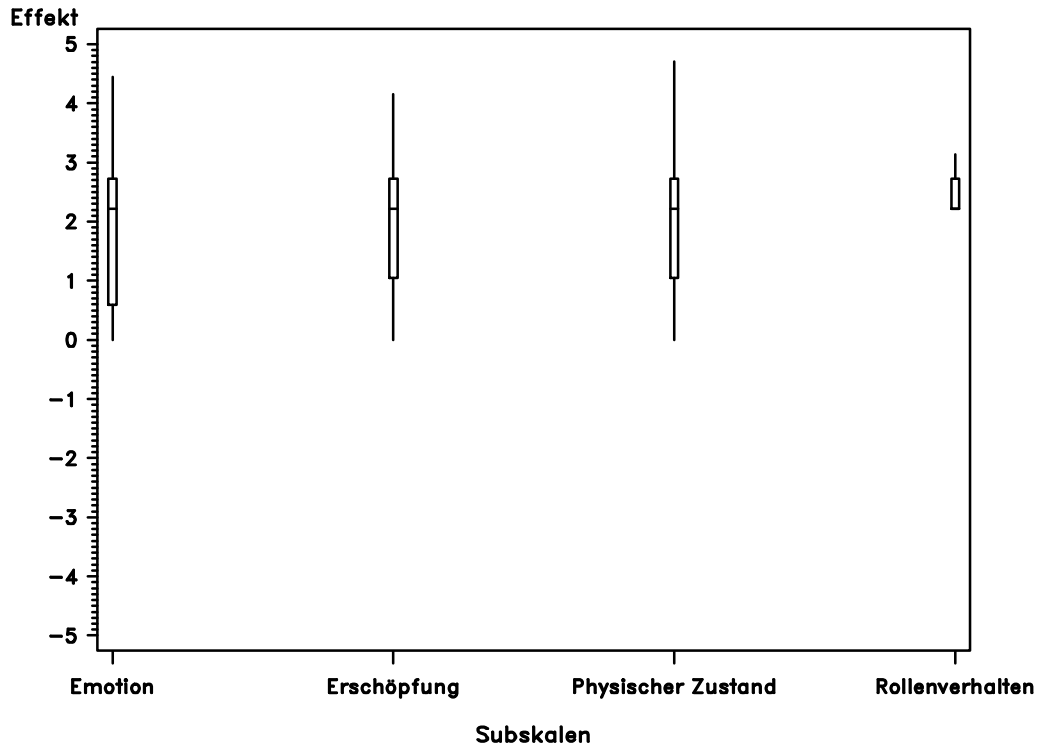

Abbildung 11.1: Boxplots der Effektstärken beim Test auf Signifikanz der Veränderung in wenigstens einer Subskala.

Zusammenhänge zwischen gemessenen Effektgrößen und anderen Variablen

Im Folgenden wird der Zusammenhang zwischen den Effektgrößen der einzelnen Traits und bestimmten anderen Variablen untersucht. Zunächst interessiert uns dabei die Frage des Zusammenhangs zwischen dem Ausgang der mehrdimensionalen Personenfittests und den Effektgrößen der einzelnen Traits. Hierdurch wollen wir herausfinden, ob alle Traits auf die Ergebnisse der Personenfittests einwirken, oder nicht.

Zu diesem Zweck setzen wir die Effektstärken derjenigen Personen auf den Wert 0, bei denen kein Personenfittest durchgeführt wurde. Weiterhin teilen wir die Stichprobe in zwei Personengruppen ein: in die Gruppe derjenigen, bei denen der Personenfittest bei Signifikanzniveau 0.10 zur Ablehnung der Nullhypothese führt, und derjenigen Personen, bei denen dies nicht der Fall ist.

Die mittleren Effektstärken dieser Gruppen wurden mit SAS/PROC SUMMARY berechnet und sind in Tabelle 11.1 ausgegeben. Wie man sieht, bestehen deutliche Unterschiede zwischen den beiden Gruppen, und zwar bei allen Traits. Dies kann durch eine Diskriminanzanalyse abgesichert werden, mittels derer untersucht wird, welche Variablen zur Trennung der beiden Gruppen verwendet werden können. Es wurde PROC DISCRIM mit nichtparametrischer Dichteschätzung und quadratischer Diskriminanzfunktion verwendet. Die Ergebnisse der Diskriminanzanalyse sind:

H_0 abgelehnt?	Anzahl	Physis	Rollen-verhalten	Emotion	Erschöpfung
Nein	160	0.394	0.357	0.606	0.518
Ja	52	2.409	1.86	1.67	1.89
Gesamt	212	0.888	0.725	0.867	0.856

Tabelle 11.1: Mittlere Effektgrößen pro Trait in Abhängigkeit vom Ausgang der Personenfittests für $\alpha = 0.1$.

- Bei jeder der vier Effektgrößen kann mittels ANOVA nachgewiesen werden, dass sich die Mittelwerte der Effektgrößen in o.g. Gruppen zum Signifikanzniveau $\alpha = 0.0001$ signifikant voneinander unterscheiden.

- Der zugehörige multivariate Test lehnt die Nullhypothese

  > H_0: Für jede der vier Effektgrößen unterscheiden sich die Mittelwerte nicht bezüglich einer „Gruppierung nach Ablehnung“

  zum Signifikanzniveau $\alpha = 0.0001$ ab. Wilks Lambda nimmt hierbei den Wert 0.370 an.

- Die zugehörige Diskriminanzfunktion ordnet 49 von 52 Personen, bei denen es im Personenfittest zu einer Ablehnung der Nullhypothese gekommen ist, der richtigen Gruppe zu. Ebenso werden 140 von 160 Personen, bei denen kein Test durchgeführt werden konnte oder keine Ablehnung erfolgte, richtig zugeordnet. Dies zeigt, dass die Effektgrößen die Einteilung nach Personenfittest gut widerspiegeln.

Weiterhin untersuchen wir den Zusammenhang zwischen den Effektgrößen und weiteren Variablen, nämlich dem Geschlecht der Versuchsperson und der Art der Behandlung. Bei der Art der Behandlung wird zwischen „ambulant“ und „stationär“ unterschieden. Die Überprüfung der multivariaten Nullhypothese

> H_0: Bei allen 4 untersuchten Effektgrößen unterscheiden sich die Gruppenmittelwerte bezüglich der Einteilung nach Geschlecht und Behandlungsart nicht

mittels multivariater Varianzanalyse und PROC GLM führt zu folgendem Ergebnis: Es sind keine zum Signifikanzniveau $\alpha = 0.05$ signifikanten Effekte bezüglich Geschlecht oder Behandlungsart festzustellen. Ebensowenig kann ein Interaktionseffekt nachgewiesen werden. Für die Variable Geschlecht nimmt die Testgröße Wilks Lambda den Wert 0.983 an, für die Variable Behandlungsart den Wert 0.989. Der Test für den Interaktionseffekt führt zu einem Wilks Lambda-Wert von 0.995. Man kann daher keinen Einfluss von Behandlungsart oder Geschlecht auf die Stärke der Veränderung nachweisen.

Schließlich wollen wir untersuchen, ob ein Zusammenhang zwischen Behandlungslänge (= Zeitraum zwischen den Befragungsterminen) und Ablehnung der Nullhypothese im

Personenfittest existiert. Dies untersuchen wir mit Hilfe eines Wilcoxon-Tests. Das Ablehnungsergebnis der Personenfittests wurde dabei folgendermaßen codiert:

- 1 falls die Nullhypothese abgelehnt wurde
- 0 sonst (also: wenn kein Test durchgeführt werden konnte oder die Nullhypothese nicht abgelehnt wurde).

Auch hier kann kein signifikanter Effekt festgestellt werden. Die (standardnormalverteilte Form der) Testgröße des Wilcoxon-Tests nimmt den Wert 0.582 an, was nicht zu einer Ablehnung für das Signifikanzniveau $\alpha = 0.05$ führt.

11.3 Konstante Veränderung bei allen Schwellen

Theorie

In diesem Abschnitt untersuchen wir, ob im LPCM für alle Itemschwellen der gleiche Veränderungsparameter angenommen werden kann. Die Tests für mehrdimensionalen Personenfit aus Kapitel 6.3 wurden zwar für dichotome Items hergeleitet, können aber, ähnlich wie in Kapitel 6.1 dargestellt, auf den Fall polytomer Items übertragen werden.

Die in Kap. 6.3 vorgestellten Testverfahren für Kontrasthypothesen können darüber hinaus verwendet werden, um zu überprüfen, ob bei einer Versuchsperson ein für alle Schwellen gleicher Veränderungsparameter angenommen werden kann. Ausgangspunkt dazu ist das in Kapitel 6.1 vorgestellte allgemeine Veränderungsmodell. Der Personenfittest in diesem Modell ist ein Binomialtest, falls nur ein Trait vorliegt. Grundlage dieses Tests ist, dass suffiziente Statistiken $x_{vki\cdot}$ für die Schwellen $1, \ldots, j, \ldots, J$ der Items $1, \ldots, i, \ldots I$ vorgegeben sind. Nach Vorgabe dieser suffizienten Statistiken sind die Antwortvektoren $(c_{v111}, c_{v112}), \ldots, (c_{vIJ1}, c_{vIJ2})$ voneinander unabhängige Zufallsvektoren.

Untersucht werden soll jetzt, ob ein schwellenunabhängiger Veränderungsparameter vorliegt. Dazu definieren wir das folgende LPCM:

$$P(X_{vi2} = h) = \frac{exp\left(h\theta_v - \sum_{j \leq h} \tau_{ij} + \sum_{j \leq h} \delta_{g2j}\right)}{\sum_l exp\left(l\theta_v - \sum_{j \leq l} \tau_{ij} + \sum_{j \leq l} \delta_{g2j}\right)} \quad . \tag{11.2}$$

Getestet wird folgendes Hypothesenpaar:

$$H_0 : \delta_{1g2} = \ldots = \delta_{Jg2} = \delta_{g2}$$

gegen

$$H_A : \delta_{kg2} \neq \delta_{lg2}$$

für mindestens ein Paar (k, l),$1 \leq k, l \leq J$. Aufgrund der Unabhängigkeit der Antwortvektoren voneinander können die in Kapitel 6.3 vorgestellten Tests auch für Gruppen von Itemschwellen verwendet werden. So können die Itemschwellen z.B. nach dem Parameter j $1 \leq j \leq J$ gruppiert werden, indem eine Gruppe aus allen Schwellen mit gleichem Indexparameter j besteht.

Realisierung

Einen Test dieser Art führen wir für die vierstufigen Items des EORTC QLQ-C30-Fragebogens durch, also für die Items 1-28. Realisiert wird dieser Test mit Hilfe der SAS-Prozedur PROC FREQ. Verwendet wird der exakte Freeman-Halton-Test. Zunächst werden Indikatorvariablen für die Schwellen der verwendeten Items berechnet. Diese Indikatoren werden anschließend gruppiert: Alle Schwellen für die Stufe 1 werden in die erste Gruppe eingeordnet, usw. Somit liegen drei Gruppen von Schwellenindikatoren mit jeweils 28 Elementen vor. Im nächsten Schritt werden dann die in 6.3 beschriebenen Tests an den Schwellenparametern mit Hilfe der obigen Gruppierung durchgeführt.

Aus unterschiedlichen Gründen kann der Personenfittest nicht an allen 296 Versuchspersonen durchgeführt werden:

- 96 Personen haben mindestens ein Item nicht beantwortet.
- Bei 42 Personen treten Items, die in genau einem Messzeitpunkt gelöst wurden, nur in einer Stufe auf. Bei diesen ist ein Vergleich zwischen unterschiedlichen Stufen nicht möglich.
- Bei 43 Personen tritt folgende Situation auf: Alle Items, die in genau einem Messzeitpunkt gelöst werden, werden im gleichen Messzeitpunkt t_1 gelöst. In diesem Extremfall berechnet PROC FREQ keine Kontingenztafel, da hier nur eine degenerierte Tabelle vorliegt. Diese Fälle werden daher ebenfalls nicht für die Personenfittests verwendet.
- Bei 28 Personen treten die beiden gerade genannten Phänomene gleichzeitig auf. Daher werden die Personenfittests aus den beiden gerade genannten Gründen bei 57 Versuchspersonen nicht durchgeführt.

Die Personenfittests werden somit für 143 Personen durchgeführt. In Tabelle 11.2 treten 12 Versuchspersonen mit einem P-Wert von höchsten 0.1 auf: Insgesamt kann also bei 12 Versuchspersonen die Nullhypothese des Tests auf Gleichheit der Veränderungsparameter abgelehnt werden. Dies genügt allerdings nicht, um ein globales Signifikanzniveau von 0.05 einzuhalten:

- Bei der Überprüfung des globalen Signifikanzniveaus mit der Binomialtestmethode ergibt sich Folgendes: Wenn als Signifikanzniveau der Einzeltests $\alpha_E = 0.1$ gewählt

Id	P-Wert
22	0.0344
75	0.0677
91	0.0512
154	0.0577
158	0.0440
203	0.0061
210	0.0750
224	0.0242
278	0.0077
284	0.0048
287	0.0047
292	0.0429

Tabelle 11.2: Versuchspersonen, bei denen die Nullhypothese für $\alpha = 0.1$ abgelehnt werden kann.

wird, ist die Zahl von 12 Ablehnungen bei 143 durchgeführten Tests nicht signifikant zum Niveau $\alpha_{Gesamt} = 0.1$. Die Wahrscheinlichkeit von mehr als 12 Ablehnungen beträgt 0.904.

- Der Test nach der Adding of Logs-Methode ermöglicht ebenfalls keine Ablehnung der Gesamtnullhypothese. Die Testgröße weist in diesem Fall den Wert 218.2 auf. Der P-Wert aus der χ^2-Verteilung mit 286 Freiheitsgraden nimmt in diesem Fall den Wert 0.999 an.

Da die Gesamtnullhypothese nicht abgelehnt werden kann, muss angenommen werden, dass die Ablehnungen der Einzelnullhypothesen nur zufällig zustande gekommen sind. Man kann daher weiter annehmen, dass bei allen Schwellen der gleiche Veränderungswert auftritt.

Kapitel 12

Diskussion

Im ersten Abschnitt dieses Kapitels werden die Ergebnisse der vorhergegangen Kapitel zusammengefasst und die Einsatzmöglichkeiten der vorgestellten Methoden diskutiert. Darauf folgt eine Diskussion der Ergebnisse der Teststärkenberechnung sowie der Auswirkungen des Ausgangswertgesetzes auf die Teststärke der Personenfittests. Schließlich werden die Einsatzmöglichkeiten für die in dieser Arbeit vorgeschlagenen Methoden in der Psychometrie zusammengefasst.

12.1 Zusammenfassung: Einsatz von Personenfittests und Effektstärken in der Veränderungsmessung

Inhalt dieser Arbeit ist die Vorstellung von Methoden zur Optimierung der Veränderungsmessung mit dem LLTM. Es werden Methoden entwickelt, die als Ergänzung zu den klassischen Signifikanztests für das LLTM und daraus abgeleiteten Modellen verwendet werden können. Kapitel 3 stellt auf dem Mantel-Haenszel-Test/dem Logitmodell aufbauende Methoden vor, die zur Identifikation von Fehlspezifikation (d.h. Subpopulationen mit – gegenüber dem angenommenen Modell – abweichenden Veränderungswirkungen) dienen können. Diese Methoden überprüfen die Gültigkeit eines vorgegebenen Veränderungstraits in einer Subpopulation und besitzen damit den gleichen Einsatzbereich wie „konventionelle“ Testverfahren für das LLTM.

Die durch [Fischer, 1995b], [Fischer, 1995a] favorisierten Likelihood-Ratio-Tests, bzw. die von [Bechger et al., 2002] favorisierten Lagrange-Multiplier-Tests benötigen jedoch die Parameterschätzungen des LLTMs zur Berechnung der jeweiligen Testgröße. Dagegen sind die (von uns entwickelten) auf dem Mantel-Haenszel-Test bzw. den Logit-Modellen aufbauenden Verfahren unabhängig von den Parameterschätzungen des LLTMs. Wie durch [Klein, 1999] bzw. [Baker, 1993] dargestellt, können Fehlspezifikationen zu deutlich verzerrten Parameterschätzungen führen. Es erscheint daher sehr fragwürdig, einen die Gültigkeit einer Teilmenge der geschätzten Parameter überprüfenden Signifikanztest

auf verzerrte Schätzwerte aufzubauen. Für die Entdeckung von Fehlspezifikation ist die in Kapitel 3 vorgestellte Methodik daher als vorteilhaft anzusehen.

Ein völlig anderer Zugang zur Erkennung von Fehlspezifikation wird durch die Kapitel 4 bis Kapitel 8 verfolgt. Wir entwickeln dort Methoden, die die Abweichungen einzelner Versuchspersonen von einem angenommenen LLTM analysieren. Bei den in Kapitel 4 bis Kapitel 7 Verfahren ist eine Kenntnis der Itemparameter nicht notwendig, da diese durch die Verwendung suffizienter Statistiken herausgerechnet werden können. Bei diesen Verfahren besitzt die Testgröße stets eine Binomial- bzw. (bei mehr als zwei Zeitpunkten, oder der gleichzeitigen Untersuchung mehrerer Traits) eine Multinomialverteilung. Im Folgenden bezeichnen wir diese Verfahren daher als Binomialtests.

Kapitel 8 behandelt Verfahren, bei denen die Schwierigkeit der Items im ersten Messzeitpunkt bekannt sein muss. Diese bieten Vorteile, wenn das Phänomen der Erinnerung an den ersten Messzeitpunkt auftritt: Die in Kapitel 4 bis Kapitel 7 beschriebenen Methoden konstruieren die Testgröße nur aus solchen Items, bei denen unterschiedliche Antworten an den beiden untersuchten Messzeitpunkten vorliegen, während die in Kapitel 8 beschriebenen Verfahren stets alle Items verwenden. Falls sich Versuchspersonen korrekt an den ersten Messzeitpunkt erinnern, verschlechtert dies bei den auf die Binomialverteilung aufbauenden Verfahren die Trennschärfe der Tests.

Andererseits besitzen die auf der Binomialverteilung aufbauenden Verfahren die Vorteile, dass keine Vorkenntnisse über Nuisance-Parameter bestehen müssen, und dass gut anwendbare Effektgrößen konstruiert werden können. Darüber hinaus lassen sich Binomialtestverfahren mit konventionellen Statistikpaketen leicht anwenden. Die in dieser Arbeit verwendeten Alternativhypothesen werden in konventionellen Quellen über die Teststärke eines Signifikanztests (vgl. z.B. [Cohen, 1969]) bisher nicht berücksichtigt. Aus diesen Gründen werden mittels einer Simulation sowie mit Teststärkeberechnungen die Eigenschaften der Binomialverteilungstests gegenüber bestimmten Alternativhypothesen untersucht.

Wie sich in Kapitel 9 und 10 zeigt, besitzen diese Binomialverteilungstests bei weniger als 30 verwendeten Items bzw. vielen an beiden Zeitpunkten gleich beantworteten Items nur eine geringe Teststärke. Dies korrespondiert mit der Literatur über die Validitätsverbesserung durch klassische Personenfitmaße (wie l_z, vgl. [Schmitt et al., 1993], [Meijer, 1997], [Meijer, 1998], [Schmitt et al., 1999]), wo keine wesentliche Verbesserung der Validität durch Personenfituntersuchungen festgestellt wurde. Die genannten Autoren begründen dies u.a. mit der geringen Teststärke von Personenfitmaßen. Die geringe Teststärke der klassischen Verfahren zur Personenfitmessung ist vermutlich auch auf die schlechte Anpassungsgüte der approximativen Verteilung der Personenfitindizes bei niedrigen Testlängen zurückzuführen (vgl. z.B. [Snijders, 2001], [Nering, 1995]). Dieser Grund für die gefundenen niedrigen Teststärken liegt bei den in dieser Arbeit vorgeschlagenen Personenfittests nicht vor, da hier stets die exakten Verteilungen verwendet wurden.

[Meijer et al., 1994] empfiehlt, Personenfituntersuchungen eher als explorative Untersuchungen durchzuführen:

> In general, person-fit analysis will be used as an exploratory technique to find respondents who behave unexpectedly on the basis of an IRT model...

Nach unserer Auffassung trifft dies auch auf die in dieser Arbeit vorgestellten Personenfittests zu. Ziel für die hier vorgestellten Verfahren ist die Entdeckung von Fehlspezifikation bei LLTMs in der Veränderungsmessung. Durch die geringe Teststärke der Personenfittests werden tatsächliche Fehlspezifikationen nur in einem schwachen Ausmaß erkannt. Sinnvoller erscheint uns daher der Einsatz der in Kapitel 7 vorgestellten Effektgrößen zu sein. Mit Hilfe dieser Effektgrößen sind quantitative Aussagen über die Abweichung einer Versuchsperson von einem vorgegebenen Modell möglich, sowie Aussagen über die durchschnittliche Abweichung innerhalb einer Subpopulation von dem vorgegebenen Modell.

Interessant ist hierbei die Parallele zur Metaanalyse. Da diese ebenfalls das Problem der möglichst sinnvollen Zusammenfassung einzelner Testgrößen behandelt, zitieren wir im Folgenden aus den Arbeiten von [Fricke and Treinies, 1985] sowie [Sedlmeier, 1996], die sich eingehender mit den Vor- und Nachteilen der Metaanalyse beschäftigen. Gemäß [Fricke and Treinies, 1985], Kap. 5.1. ist in einer Metaanalyse das reine Auszählen von Signifikanzen nur bei hohen Effektstärken und großen Stichprobengrößen der Einzelstudien effektiv. Fricke/Treinies kommen zu dem Ergebnis, dass der

> ... Einsatz dieser Methoden deshalb nur dann angebracht ist, wenn die mitunter zu spärlichen statistischen Angaben in den Primärstudien keine andere Methodenwahl zulassen.

Die Autoren setzen allerdings den Einsatz einer x%-Auswahlregel voraus, wie z.B. „Man entscheide sich für das Vorliegen eines Effektes, wenn mehr als x% der Einzelstudien zu einer signifikanten Ablehnung der Nullhypothese geführt haben“. Das von uns in Kapitel 7.1 entwickelte Verfahren ist wesentlich trennschärfer als eine solche x%-Auswahlregel, doch verliert man auch bei diesem Verfahren Information durch die Einteilung der Versuchspersonen in nur 2 Klassen.

Auch das Adding-Of-Logs-Verfahren und andere Methoden zur Zusammenfassung von P-Werten (vgl. Kapitel 7.2 in der vorliegenden Arbeit) werden von [Fricke and Treinies, 1985] als suboptimal beurteilt. Durch solche Verfahren kann zwar ein von 0 verschiedener Gesamteffekt entdeckt werden, andererseits sind jedoch Aussagen über die Höhe des Durchschnittseffektes bzw. die Streuung der Effekte nicht möglich. Gerade solche Aussagen sind jedoch hilfreich, um das Verhalten von Abweichungen zum postulierten Modell zu beschreiben (vgl. z.B. Abb. 11.1).

Gegen das einfache Aggregieren von P-Werten spricht nach Meinung des Autors ebenfalls, dass die Neyman-Pearsonsche Testtheorie nur die Aussageformen signifikant und nicht signifikant kennt. Eine zum Niveau $\alpha = 0.05$ signifikante Aussage hat gemäß der Neyman-Pearsonschen Testtheorie nicht weniger Gewicht als eine zum Niveau $\alpha = 0.01$ signifikante Aussage. Ein Vergleich von P-Werten, wie er mittels der Adding-Of-Logs-Methode erfolgt, ist in der Neyman-Pearsonschen Testtheorie nicht erlaubt. Ähnliche Aussagen macht [Sedlmeier, 1996], der hierbei zwischen der Fisherschen Signifikanztesttheorie und der Neyman-Pearsonschen Theorie unterscheidet.

Somit stellt unser Vorschlag zur Verwendung von Effektgrößen für den Personenfit und deren Analyse einen sinnvollen Ansatz zur Untersuchung auf Personenfit im LLTM dar.

12.2 Diskussion der Teststärkeberechnung

Teststärke und Zahl der eingehenden Items

Verwendet man das Binomialtestverfahren, werden akzeptable Teststärken erst bei einer großen Anzahl eingehender Items erreicht. Bei einer zu entdeckenden Effektstärke von 0.3 genügen ca. 30 Items, um eine größere Teststärke als 0.9 zu erreichen. Bei Effekten der Effektstärke 0.1 kann dagegen erst bei mehr als 100 Items eine Teststärke von 0.9 erreicht werden. In die eigentlichen Parameterschätzungen umgerechnet bedeutet eine Effektstärke von 0.1 z.B. einen Anstieg von $\delta = 0$ auf $\delta = 0.4$, und eine Effektstärke von 0.3 z.B. einen Anstieg von $\delta = 0$ auf $\delta = 1.2$.

In den Binomialtest gehen nur solche Items ein, die an beiden Messzeitpunkten unterschiedlich beantwortet werden. Da dies in der Regel nicht bei allen verwendeten Items der Fall sein wird, benötigt man in der Praxis noch wesentlich mehr Items als in dem oben genannten Beispiel.

Gleiche Antworten an beiden Messzeitpunkten können aus verschiedenen Gründen auftreten:

- zufällig, bzw.
- die Fähigkeit der Versuchsperson überschreitet deutlich die Itemschwierigkeit, bzw.
- es herrscht ein starker Zusammenhang zwischen dem Antwortverhalten an den untersuchten Messzeitpunkten (= z.B. „Erinnern an die Antwort im ersten Messzeitpunkt“).

Durch diese 3 Phänomene kann die Teststärke in der Praxis deutlich vermindert werden.

Um Personenfittests sinnvoll einzusetzen, sollten daher folgende Voraussetzungen gegeben sein:

- Hohe Zahl verwendeter Items (resp. Antwortschwellen).
- Maßnahmen zur Vermeidung eines Zusammenhangs zwischen den Antworten im ersten und zweiten Messzeitpunkt: Z.B. ausreichend langer Zeitraum zwischen den beiden Zeitpunkten oder die Verwendung von Items, bei denen das Phänomen „Erinnerung“ nicht auftreten kann.
- Schließlich ist in Fällen, bei denen die gerade genannten Punkte nicht anzuwenden sind, ein Versuchsplan zu verwenden, der die Beantwortung unterschiedlicher Items an den beiden Antwortzeitpunkten vorsieht. Ein Messmodell für diesen Fall wurde in [Fischer, 1995c] beschrieben. Als Personenfittests müssen in diesem Fall die Tests aus Kapitel 8.1 verwendet werden.

Zusammenhang zwischen Veränderung und Fähigkeit im ersten Messzeitpunkt

Eine Verminderung der Teststärke kann auch durch einen Zusammenhang zwischen Veränderung und Fähigkeit bewirkt werden. Dies geschieht immer dann, wenn der Zusammenhang „gleichgerichtet“ ist, d.h. wenn die Veränderung eine hohe Fähigkeit noch verstärkt bzw. eine niedrige Fähigkeit weiter absenkt. Diese Eigenschaft der Personenfittests korrespondiert mit dem Ausgangswertgesetz von [Wilder, 1931] (vgl. dazu auch [Krause, 1997]) und stellt wohl ein Problem jeglicher Veränderungsmessung dar.

Generell ist im LLTM die Ausgangswertabhängigkeit unkritisch: Da die Schätzung der Veränderungsparameter pro Gruppe erfolgt, beruht die Schätzung der Veränderungsparameter vor allem auf Personen mit mittleren Fähigkeiten. Personen mit hoher Fähigkeit besitzen einen großen Anteil von Items, die an beiden Zeitpunkten gelöst wurden. Bezüglich einer weiteren Erhöhung der Fähigkeit ist bei diesem Personenkreis nur wenig Information enthalten. Ein ähnlicher Zusammenhang gilt bei Personen mit geringer Fähigkeit, wenn die Fähigkeit weiter erniedrigt wird.

Dies gilt allerdings nur für die Parameterschätzung. Aufgrund der Form der logistischen Funktion steigt im LLTM die Itemantwortwahrscheinlichkeit nur in geringem Maß an, falls ein hoher Fähigkeitsparameter weiter erhöht wird. Durch die Wahl der logistischen Funktion als Itemantwortfunktion kann das LLTM daher gut auf die Ausgangswertabhängigkeit reagieren.

Dies gilt jedoch nicht mehr, wenn man die Antwortvektoren einzelner Personen beobachtet, wie das bei Personenfittests geschieht. Bei Personen mit extremen Fähigkeiten und gleichgerichteter Veränderung liegt eine niedrige Zahl ungleich beantworteter Items vor

und somit eine niedrige Teststärke.

Dieses Problem existiert übrigens auch bei den Tests aus Kapitel 8.1. Bei einer hohen Zahl gleich beantworteter Items existiert nur eine niedrige Zahl möglicher Antwortmuster mit vorgegebenem Summenscore, deren Antwortwahrscheinlichkeiten zudem stets sehr ähnlich sind. Außerdem ändern sich die Wahrscheinlichkeiten solcher Antwortvektoren nur in geringem Maße, wenn der Veränderungsparameter einen anderen Wert annimmt. Dies folgt, wie man leicht erkennen kann, aus der Voraussetzung einer logistischen Itemantwortfunktion.

„Nicht gleichgerichtete“ Zusammenhänge zwischen Fähigkeit und Veränderung können dagegen mit Personenfittests hervorragend aufgedeckt werden. Somit können Zusammenhänge zwischen Fähigkeit und Veränderung auch durch Beobachten der Effektgrößen von Personenfittests entdeckt werden: Niedrige Effektstärken bei mittleren und hohen Fähigkeiten, sowie hohe Effektstärken bei niedrigen Fähigkeiten deuten auf einen „nicht gleichgerichteten“ Zusammenhang zwischen Fähigkeit und Veränderung hin. Dies kann grafisch leicht überpüft werden.

Eignung von Personenfittests zur Entdeckung von Bias bezüglich der Veränderung

Als Bias bezüglich der Veränderung bezeichnen wir das Phänomen, dass in einer Teilmenge von Items in einer Subpopulation ein anderer Wert des Veränderungsparameters angenommen werden muss als bei den restlichen Items. [Ponocny, 2000] schlägt den Einsatz von Personenfittests zur Erkennung von Item Bias vor. Zur Entdeckung dieser Art von Verzerrung eignen sich Personenfittests nur dann, wenn man die Items schon kennt, bei denen diese Verzerrung auftritt. Wie in Kapitel 10.3 dargelegt, kann man durch den naiven Einsatz des Binomialtests solche Verzerrungen nicht erkennen, wenn der Mittelwert der durch den Bias beeinflussten Parameterwerte gleich dem wahren Veränderungsparameter ist. Auch die in Kapitel 9 durchgeführte Simulation lässt auf eine äußerst niedrige Teststärke der Personenfittests schließen, falls Abweichungen zur Nullhypothese nur bei einigen wenigen Items auftreten und/oder kein Test Bias vorliegt. Die Verwendung des Binomialtests zur Entdeckung kann daher nicht empfohlen werden. Abhilfe bieten mehrdimensionale Tests oder die Beschränkung des auf diejenigen Items, bei denen ein Bias vermutet wird.

Im Vergleich zu den in Kapitel 3 vorgestellten Mantel-Haenszel-Tests auf Bias bezüglich der Veränderung ist die Verwendung von Personenfittests außerdem unhandlich: Man muss für jede Person einen Test durchführen, während bei der Methode aus Kapitel 3 nur ein Test pro verdächtigem Item nötig ist. Zudem vermindert eine falsche Auswahl potenziell verzerrter Items die Teststärke der Personenfittests.

Trotzdem erscheint in manchen Fällen die Verwendung der Personenfittests zur Entdeckung von Bias bezüglich der Veränderung vernünftig: Falls sehr viele Items benutzt werden, wächst die Zahl der vom Mantel-Haenszel-Test aus Kap. 3 verwendeten Zellen stark an. Oftmals wird dies dazu führen, dass die beobachtete Stichprobe zu klein ist, um gültige Aussagen mit einem Mantel-Haenszel-Test machen zu können. In solchen Fällen ist der Einsatz eines Personenfittests eine vernünftige Alternative. Zudem wächst die Teststärke der Personenfittests mit der Zahl der verwendeten Items.

12.3 Die Auswirkungen des Ausgangswertgesetzes

Wie schon in den vorhergehenden Kapiteln dargestellt, hängt die Teststärke der hier eingeführten Verfahren vom Ausgangswert der Versuchsperson ab. Dies ähnelt dem bekannten Phänomen der Ausgangswertabhängigkeit in der Veränderungsmessung. Dieses Phänomen beschreibt [Raykov, 1987] als eine Variante des sog. „Anfangswertproblems“: Versuchspersonen mit extremen Werten bei der ersten Messung tendieren bei der zweiten Messung eher zur Mitte. [Raykov, 1987], S. 103 unterscheidet für dieses Phänomen folgende Fälle:

- Statistische Ursachen wie z.B. die negative Korrelation zwischen Ausgangswert X_{vi1} und zugehörigem Differenzwert $X_{vi2} - X_{vi1}$,
- formale Ursachen wie z.B. Boden- und Deckeneffekte, sowie
- sachlogische Ursachen.

Wie [Raykov, 1987], S.103 anmerkt, können sachlogische und formale Ursachen oft nur schwer voneinander unterschieden werden. In unserem Fall ist nun von Decken- und Bodeneffekten und somit einer formalen Ursache auszugehen, durch die die verringerte Teststärke verursacht wird: Sei dazu mit $t \in \{1, 2\}$ und $\delta_1 = 0$ definiert:

$$p_t = \frac{\exp\left[\theta_v - \beta_i + \delta_t\right]}{1 + \exp\left[\theta_v - \beta_i + \delta_t\right]} \,. \tag{12.1}$$

p_t stellt die Wahrscheinlichkeit dar, dass ein Item in Zeitpunkt 1 resp. 2 richtig beantwortet wird. Für extreme Ausgangswerte $\theta_v - \beta_i$ und gleichgerichtete Veränderungen gilt nun $p_1 \approx p_2$, sowie $p_1 \approx 1$ oder $p_1 \approx 0$. Dies führt aber dazu, dass es bei extremen Anfangswerten häufig in beiden Zeitpunkten zu identischen Antworten kommt. Die Situation des von uns vorgeschlagenen Binomialtests lässt sich mit Hilfe einer Vierfeldertafel beschreiben:

H_{00}	H_{01}
H_{10}	H_{11}

Die Zeilen entsprechen in dieser Darstellungsweise dem ersten Messzeitpunkt, die Spalten dem zweiten Messzeitpunkt. H_{00} ist die Zahl der in beiden Zeitpunkten nicht beantworteten Items. In den Binomialtest gehen die Häufigkeiten H_{01} und H_{10} ein. Bei extremen Anfangswerten und gleichgerichteten Veränderungen sind diese Zellen schwach besetzt, während die H_{11} resp. H_{00} hohe Besetzungszahlen aufweisen. Der Binomialtest kann in diesem Fall nur die Information weniger Items verwenden und besitzt daher eine niedrige Teststärke. In unserem Fall ist daher ein Bodeneffekt zu vermerken. Somit fällt die hier aufgetretene Version des Anfangswertproblems nicht unter die Kategorie „Statistische Ursachen“.

Aus all diesem ist die Konsequenz zu ziehen, dass man die Effektgrößen der Personenfittests nur als Maß für die Abweichung von einer vorgegebenen Veränderung verwenden sollte. Als Maß für die tatsächliche Veränderung bei einer Person sind sie wegen der Ausgangswertabhängigkeit ungeeignet: Veränderungen bei hohen Fähigkeiten können mit Hilfe der vorgestellten Effektgrößen nicht gemessen werden.

12.4 Folgerungen

Insgesamt gesehen ergeben sich durch die vorliegende Arbeit folgende Konsequenzen für die Konstruktion von Messverfahren für die Veränderungsmessung:

1. Gegen die Verwendung der hier vorgestellten Signifikanztests spricht deren oft geringe Teststärke. Ähnlich wie bei den klassischen Testgrößen der Personenfitanalyse (vgl. [Schmitt et al., 1993], [Meijer, 1997], [Meijer, 1998], [Schmitt et al., 1999]) ist daher keine wesentliche Verbesserung der Validität zu erwarten. Als Alternative bietet sich hier der Einsatz von Effektstärken an.

2. Die vorgestellten Effektgrößen sollten nur zur Beurteilung von Abweichungen gegenüber anderweitig gemessenen Parameterwerten verwendet werden, nicht aber als Maß für die absolute Veränderung.

3. Die Verwendung von Effektstärken anstelle eines Signifikanztests erscheint vor allem im Hinblick auf die dadurch ausgelösten Aktionen sinnvoll. Bei der Veränderungsmessung mit dem LLTM oder einem daraus abgeleiteten Modell kann die Validität des Modells durch Veränderung der Modellparametrisierung erhöht werden. Diese Möglichkeiten existieren bei der klassischen Personenfitmessung nicht, bei der es lediglich um das Entfernen schlecht angepasster Personen aus der Stichprobe geht.

4. Die Benutzung von Effektgrößen statt Personenfittests vermeidet den durch die dichotome Testentscheidung auftretenden Informationsverlust und ermöglicht somit eine bessere Beurteilung der aufgetretenen Abweichungen. Insbesondere ist auch eine explorative Analyse der (empirischen) Verteilung der Effektstärken möglich.

12.5 Verbesserungsvorschläge für die Evaluationsstandards in der Psychotherapieforschung

Im Folgenden wird dargestellt, was die in dieser Arbeit vorgeschlagene Methodik zur Behebung der von [Metzler and Krause, 1997] angesprochenen Probleme der Zielgrößenauswahl resp. der statistischen Stichprobenplanung leisten können (vgl. auch Kap. 2.1). Die in dieser Arbeit vorgestellten Methoden erlauben eine genauere Beschreibung der Abweichung von dem postulierten Modell. Insbesondere der Einsatz von Effektgrößen erscheint sinnvoll:

Erstens wird die Stichprobenplanung bei Verwendung des LLTMs in der Veränderungsmessung erleichtert. Im Gegensatz zur reinen Anwendung von Anpassungstests können, wie in Kapitel 11.2 dargestellt, zielgerichtete Aussagen über die Homogenität und die mittlere Stärke einer Veränderung gemacht werden. Da die Verteilung der Effektstärken bekannt ist, kann dies auch (a priori) zur Planung der Stichprobengröße verwendet werden. Eine mögliche Vorgehensweise wäre z.B.:

> Bestimme die Größe der Stichprobe so, dass ein mittelerer (personenbezogener) Effekt der Größe μ mit Sicherheitsgrad γ gerade noch erkannt werden kann.

Diese Vorgehensweise weist interpretative Vorteile gegenüber dem „normalen“ Vorgehen bei der Stichprobenplanung für das LLTM auf. Im Normalfall wird im LLTM ein Likelihood-Ratio-Test zur Hypothesenprüfung verwendet. Dann würde man eine χ^2-basierte Größe für die Stichprobenplanung verwenden und die Stichprobenplanung auf die Gesamtstärke der Abweichungen von der Nullhypothese aufbauen. Falls man sich hingegen für die in dieser Arbeit vorgeschlagene Methodik entscheidet, kann man als zusätzliche Informationen die Stärke der Abweichung pro Person sowie den Anteil der abweichenden Personen in die Planung der Stichprobengröße einbeziehen: Da die personenbezogenen Effektgrößen (wie in Kapitel 7.4 dargestellt) bei Nichtgültigkeit der Nullhypothese $N(\mu; 1)$-verteilt sind, sind bei Kenntnis des wahren (Effekt-)Mittelwertes μ auch sämtliche Quantile der Verteilung der Effektgrößen bekannt. Eine Aussage über den Mittelwert μ kann daher eindeutig in eine Aussage der Form

> Bei x % der Versuchspersonen zeigt sich ein Effekt, der größer als eine Zahl h ist.

umgeformt werden. Eine solche Äquivalenz existiert nicht bei der Verwendung χ^2-basierter Größen zur Stichprobenplanung. Unser Verfahren sehen wir daher als bedeutende Erweiterung gegenüber herkömmlichen Prüfverfahren für das LLTM an.

Zweitens können die in dieser Arbeit vorgestellten Methoden auch dazu dienen, eine Brücke zwischen Einzelfall- und Gruppenmethodik zu schlagen, und somit Vorbehalte

gegenüber Gruppenuntersuchungen zu vermindern, wie sie z.B. in [Sbandi et al., 1993] dargestellt wurden. Durch die Verwendung von Personenfittests bzw. der zugehörigen Effektgrößen lassen sich individuelle Abweichungen von den mittleren Effekten dokumentieren. Verglichen mit den Standardmethoden des LLTMs ergeben sich dadurch erweiterte Darstellungsmöglichkeiten bezüglich der Wirkungsweise einer Therapie.

Drittens ermöglicht die Verwendung von Personenfit-Effektgrößen Vergleiche zwischen verschiedenen Traits hinsichtlich Stärke und Variation des Gesamteffekts, wie in Kapitel 11.2 demonstriert. Dies erfüllt die von [Hager, 1998] erhobene Forderung nach besserer Kontrolle von Wirkzusammenhängen.

Zusammenfassend lässt sich daher sagen, dass die in dieser Arbeit vorgestellten Personenfittests neben ihren Einsatzmöglichkeiten zur Entdeckung von Fehlspezifikation in Veränderungsmodellen auch zur besseren Anwendbarkeit des LLTMs bei Evaluationsuntersuchungen z.B. im Bereich der Psychotherapien beitragen.

Anhang A

Abbildungen

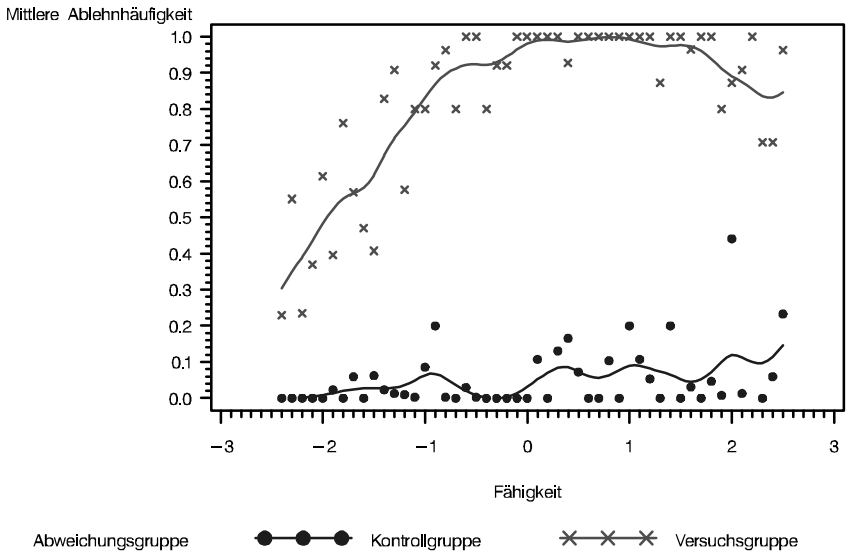

Abbildung A.1: Mittlere Ablehnwahrscheinlichkeit vs. Fähigkeit getrennt nach Zugehörigkeit zu Untersuchungsgruppe resp. Kontrollgruppe. Veränderung unabhängig von Fähigkeit. Glättung durch kubische Splinefunktion.

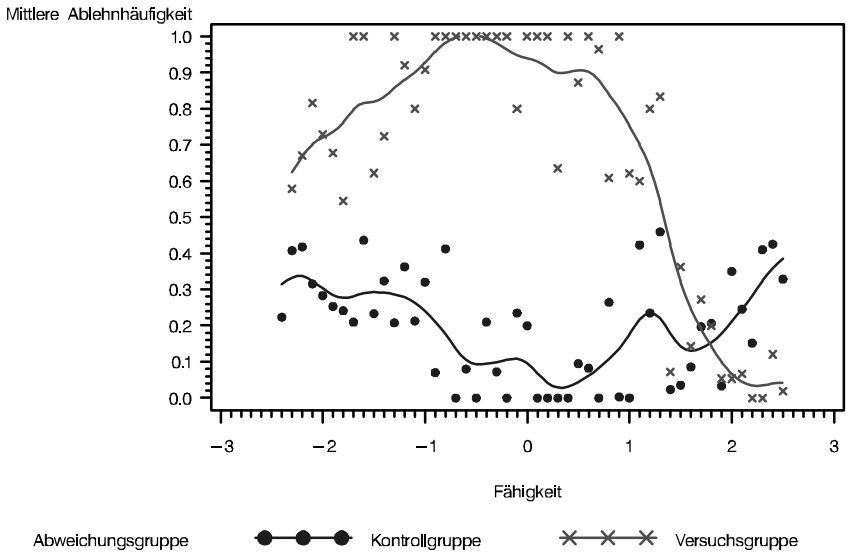

Abbildung A.2: Mittlere Ablehnwahrscheinlichkeit vs. Fähigkeit getrennt nach Zugehörigkeit zu Untersuchungsgruppe resp. Kontrollgruppe. Veränderung linear abhängig von Fähigkeit mit positivem Koeffizienten. Glättung durch kubische Splinefunktion.

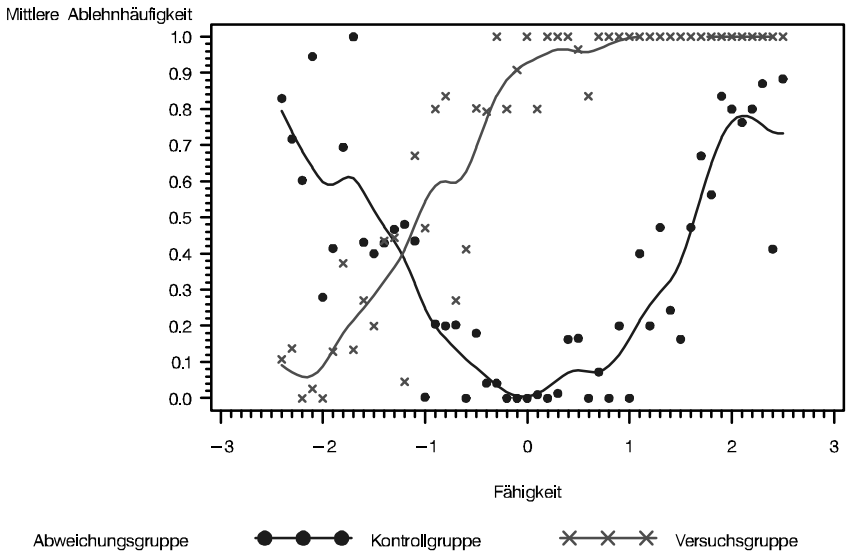

Abbildung A.3: Mittlere Ablehnwahrscheinlichkeit vs. Fähigkeit getrennt nach Zugehörigkeit zu Untersuchungsgruppe resp. Kontrollgruppe. Veränderung linear abhängig von Fähigkeit mit negativem Koeffizienten. Glättung durch kubische Splinefunktion.

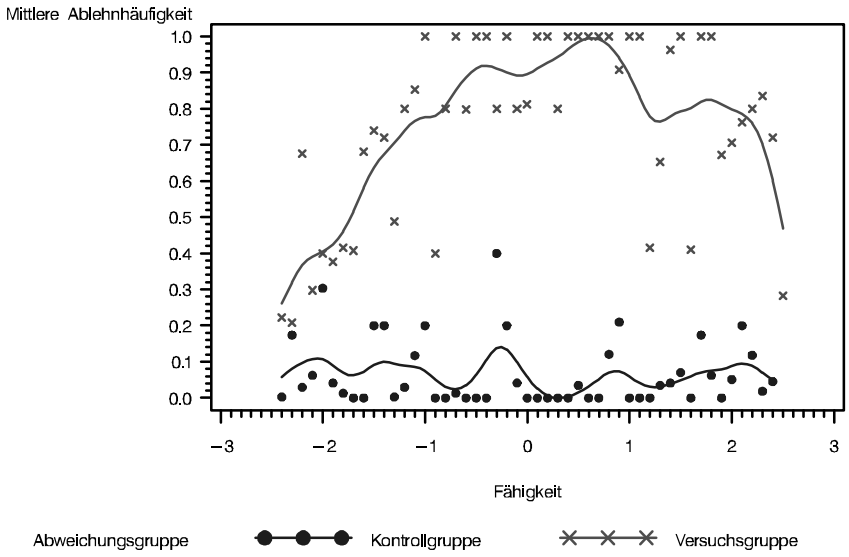

Abbildung A.4: Mittlere Ablehnwahrscheinlichkeit vs. Fähigkeit getrennt nach Zugehörigkeit zu Untersuchungsgruppe resp. Kontrollgruppe. Quadratischer Zusammenhang zwischen Veränderung und Fähigkeit. Glättung durch kubische Splinefunktion.

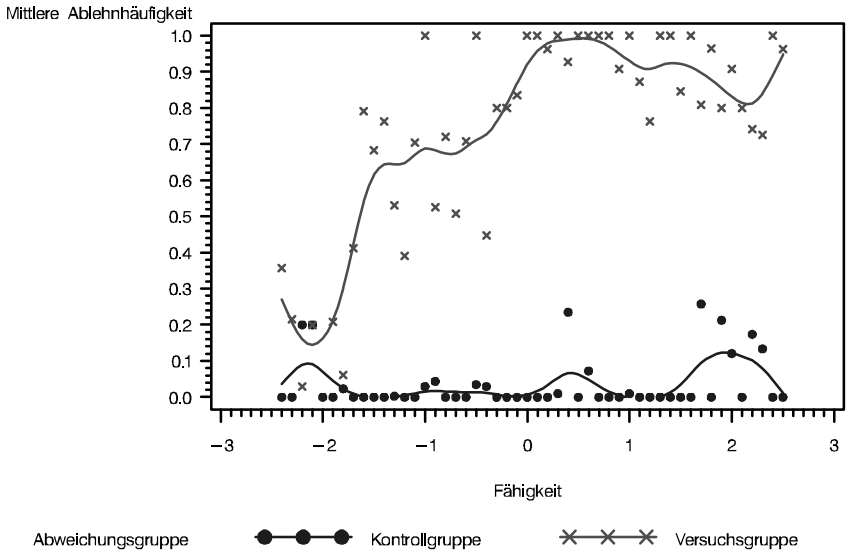

Abbildung A.5: Mittlere Ablehnwahrscheinlichkeit vs. Fähigkeit getrennt nach Zugehörigkeit zu Untersuchungsgruppe resp. Kontrollgruppe. Item Bias bei 10 Items. Glättung durch kubische Splinefunktion.

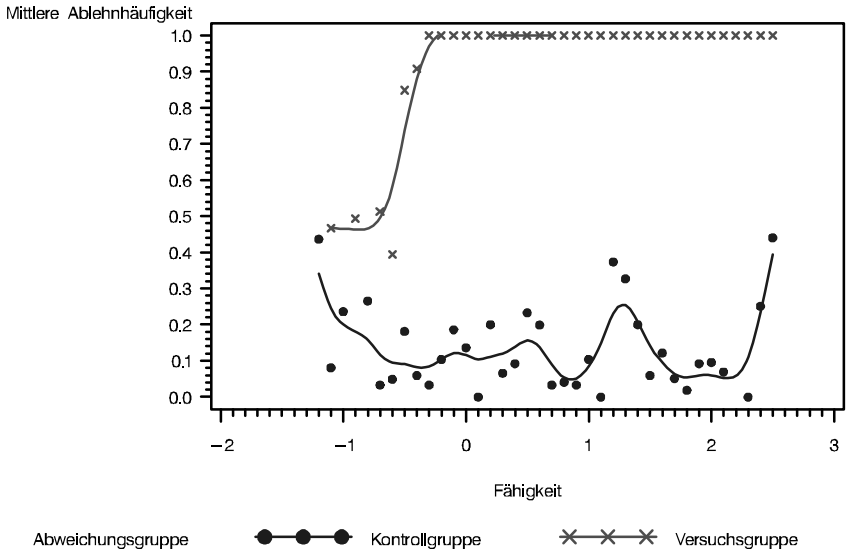

Abbildung A.6: Mittlere Ablehnwahrscheinlichkeit vs. Fähigkeit getrennt nach Zugehörigkeit zu Untersuchungsgruppe resp. Kontrollgruppe. Lineare Traitfunktion mit variierenden Trennschärfen. Glättung durch kubische Splinefunktion.

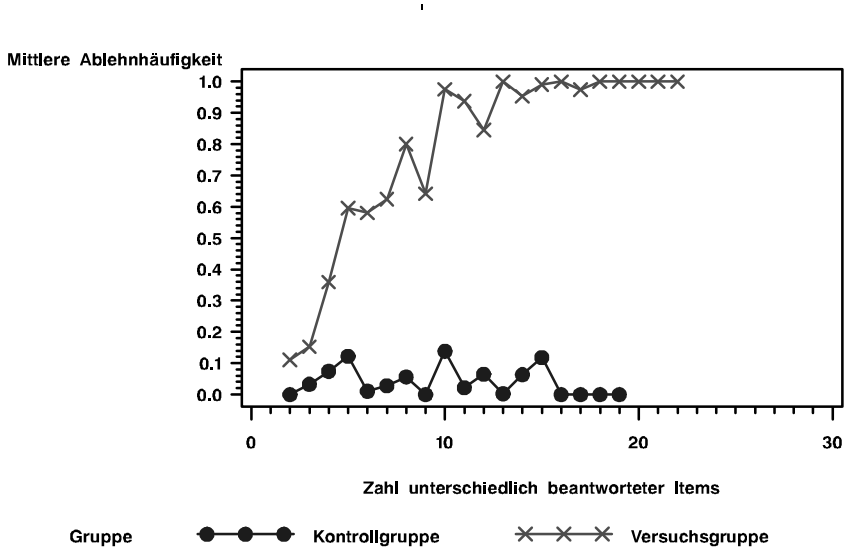

Abbildung A.7: Mittlere Ablehnwahrscheinlichkeit vs. Zahl eingehender Items getrennt nach Zugehörigkeit zu Untersuchungsgruppe resp. Kontrollgruppe. Veränderung unabhängig von Fähigkeit. Intervallförmige Nullhypothese.

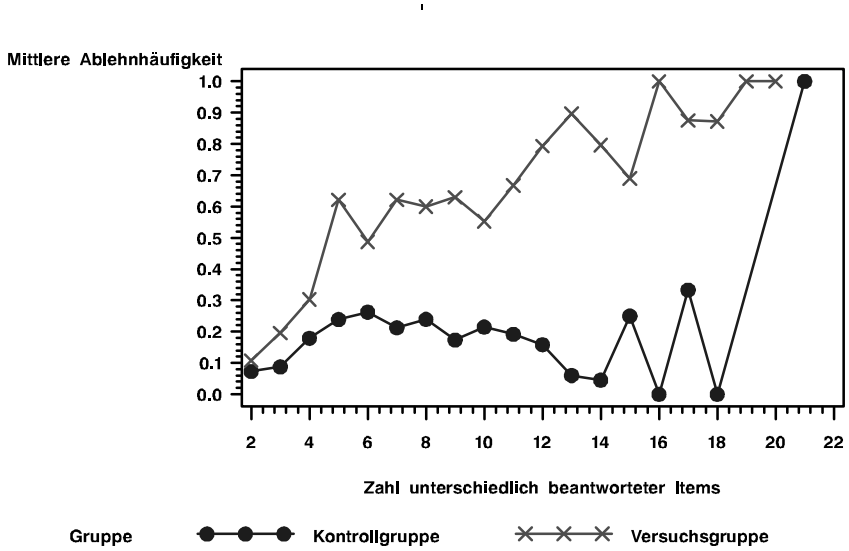

Abbildung A.8: Mittlere Ablehnwahrscheinlichkeit vs. Zahl eingehender Items getrennt nach Zugehörigkeit zu Untersuchungsgruppe resp. Kontrollgruppe. Positiver linearer Zusammenhang zwischen Veränderung und Fähigkeit. Intervallförmige Nullhypothese.

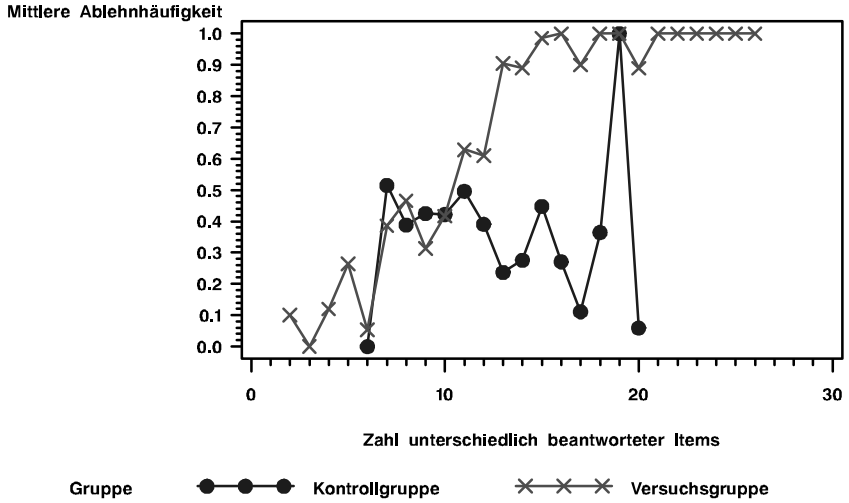

Abbildung A.9: Mittlere Ablehnwahrscheinlichkeit vs. Zahl eingehender Items getrennt nach Zugehörigkeit zu Untersuchungsgruppe resp. Kontrollgruppe. Negativer linearer Zusammenhang zwischen Veränderung und Fähigkeit. Intervallförmige Nullhypothese.

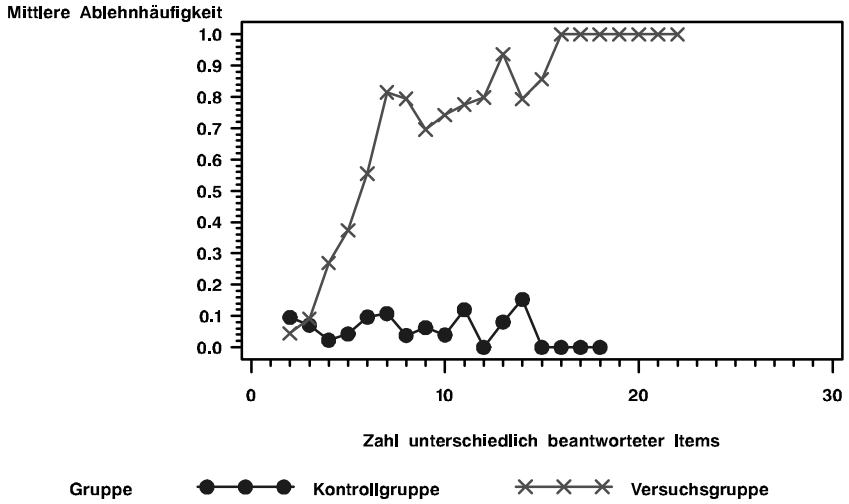

Abbildung A.10: Mittlere Ablehnwahrscheinlichkeit vs. Zahl eingehender Items getrennt nach Zugehörigkeit zu Untersuchungsgruppe resp. Kontrollgruppe. Quadratischer Zusammenhang zwischen Veränderung und Fähigkeit. Intervallförmige Nullhypothese.

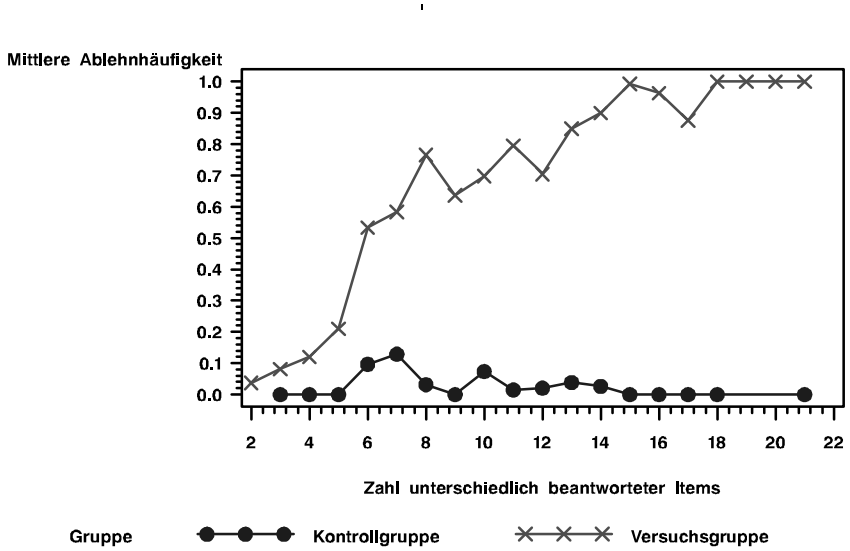

Abbildung A.11: Mittlere Ablehnwahrscheinlichkeit vs. Zahl eingehender Items getrennt nach Zugehörigkeit zu Untersuchungsgruppe resp. Kontrollgruppe. Verzerrung durch Differential Item Functioning in 10 Items. Intervallförmige Nullhypothese.

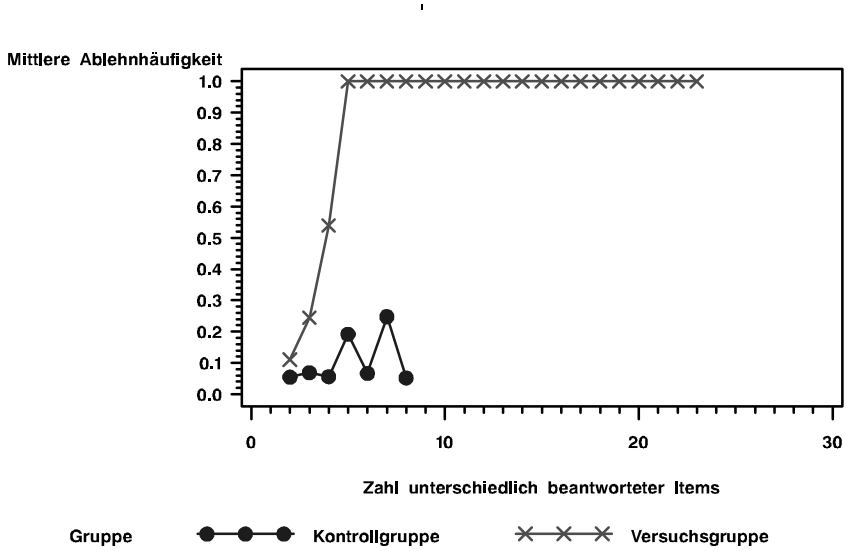

Abbildung A.12: Mittlere Ablehnwahrscheinlichkeit vs. Zahl eingehender Items getrennt nach Zugehörigkeit zu Untersuchungsgruppe resp. Kontrollgruppe. Verzerrung durch lineare Traitfunktionen mit unterschiedlichen Fähigkeiten. Intervallförmige Nullhypothese.

Anhang B

Abkürzungen

CML: Conditional Maximum Likelihood

DIF: Differential Item Functioning

EORTC QLQ-C30: European Organization for the Research and Treatment of Cancer: Quality of Life Questionaire C30

GPSD: Generalized Power Series Distributions

GLM: Generalized Linear Model

ICC: Item Characteristic Curve

IRT: Item Response Theory

KTT: Klassische Test-Theorie

LLTM: Linear Logistic Test Model

LLRA: Linear Logistic model with Relaxed Assumptions

MRM: Mixed Rasch Modell

MML: Marginal Maximum Likelihood

nMML: nonparametric Marginal Maximum Likelihood

PRF: Person Response Function

SAT: Scholastic Aptitude Test

TG: Testgröße

Literaturverzeichnis

[Aaronson et al., 1993] Aaronson, N. K. et al. (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute*, 85:365–373.

[Andersen, 1985] Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, 50:3–16.

[Andersen, 1995] Andersen, E. B. (1995). *Polytomous Rasch Models and their estimation*, pages 271–291. In [Fischer and Molenaar, 1995].

[Aydemir, 1994] Aydemir, l. (1994). *Hypothesen und Fallzahlschätzung*, pages 136–143. In [Hasford and Staib, 1994].

[Baker, 1993] Baker, F. B. (1993). Sensitivity of the linear logistic test model to misspecification of the weight matrix. *Applied Psychological Measurement*, 17:201–210.

[Bechger et al., 2002] Bechger, T. M., Verstralen, H. H. F. M., and Verhelst, N. D. (2002). Equivalent linear logistic test models. *Psychometrika*, 67:123–136.

[Bedrick, 1997] Bedrick, E. J. (1997). Approximating the conditional distribution of person fit indexes for checking the rasch model. *Psychometrika*, 62:191–199.

[Bereiter, 1963] Bereiter, C. (1963). *Some persisting dilemmas in the measurement of change*, pages 3–20. In [Harris, 1963].

[Berger, 1985] Berger, J. (1985). *Decision Theory*. Springer, Berlin Heidelberg New York.

[Cohen, 1969] Cohen, J. (1969). *Statistical Power analyses for the behavioral sciences*. Academic Press, New York.

[Cohen, 1992] Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112:155 –159.

[Cronbach and Furby, 1970] Cronbach, L. and Furby, L. (1970). How should we measure change– or should we? *Psychological Bulletin*, 74:68–80.

[Drasgow and Levine, 1986] Drasgow, F. and Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement*, 10:59–67.

[Embretson, 1991] Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, 56:495–515.

[Fahrmeir and Tutz, 1994] Fahrmeir, L. and Tutz, G. (1994). *Multivariate statistical Modelling based on generalized linear models*. Springer, Berlin Heidelberg New York.

[Falk et al., 1995] Falk, M., Becker, R., and Marohn, F. (1995). *Angewandte Statistik mit SAS*. Springer, Berlin Heidelberg New York.

[Ferrando and Chico, 2001] Ferrando, P. J. and Chico, E. (2001). Detecting dissimulation in personality testscores: a comparison between person-fit indices and detection scales. *Educational and Psychological Measurement*, 61:997–1012.

[Fischer, 1972] Fischer, G. H. (1972). A measurement model for the effect of mass-media. *Acta Psychologica*, 36:207–220.

[Fischer, 1993] Fischer, G. H. (1993). Notes on the mantel-haenszel-procedure and another chi-squared test for the assessment of dif. *Methodika*, 7:88–100.

[Fischer, 1995a] Fischer, G. H. (1995a). *Linear Logistic Models for Change*, pages 157–180. In [Fischer and Molenaar, 1995].

[Fischer, 1995b] Fischer, G. H. (1995b). *The Linear Logistic Test Model*, pages 131–155. In [Fischer and Molenaar, 1995].

[Fischer, 1995c] Fischer, G. H. (1995c). Some neglected problems in irt. *Psychometrika*, 60.

[Fischer and Molenaar, 1995] Fischer, G. H. and Molenaar, I. W., editors (1995). *Rasch Models: Foundations, Recent Developments and Applications*. Springer, Berlin Heidelberg New York.

[Fischer and Ponocny, 1994] Fischer, G. H. and Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika*, 59:177–192.

[Freeman and Halton, 1951] Freeman, G. and Halton, J. (1951). Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika*, 38:141 –149.

[Fricke and Treinies, 1985] Fricke, R. and Treinies, G. (1985). *Einführung in die Metaanalyse*. Huber, Bern (u.a.).

[Grawe, 1987] Grawe, K. (1987). *Die Effekte der Psychotherapie*, pages 515–534. Hogrefe, Göttingen.

[Grawe et al., 1994] Grawe, K., Donati, R., and Bernauer, F., editors (1994). *Psychotherapie im Wandel: von der Konfession zur Profession*. Hogrefe, Göttingen.

[Hager, 1998] Hager, W. (1998). Therapieevaluation: Begriffsbildung, Kontrolle, Randomisierung und statistische Auswertung. *Methods of Psychological Research Online*, 3:69 – 81.

[Harder, 1994] Harder, S. (1994). *Grundlagen des Wirksamkeitsnachweises: Messung von Surrogatkriterien und Surrogatendprodukten*, pages 118–127. In [Hasford and Staib, 1994].

[Harris, 1963] Harris, C. W., editor (1963). *Problems in measuring change*. The University of Wisconsin Press, Madison.

[Hasford and Staib, 1994] Hasford, J. and Staib, A. H., editors (1994). *Arzneimittelprüfungen und Good Clinical Practice*. MMV Medizin Verlag, München.

[Hoijtink, 1995] Hoijtink, H. (1995). *Linear and repeated measures models for the person parameters*, pages 203–214. In [Fischer and Molenaar, 1995].

[Holland, 1990] Holland, P. W. (1990). On the sampling theory foundations of item response theory. *Psychometrika*, 55:577–601.

[Holland and Thayer, 1988] Holland, P. W. and Thayer, D. T. (1988). *Differential Item performance and the Mantel Haenszel Procedure*, pages 129–145. Lawrence Erlbaum, Hillsdale, NJ.

[Hornke and Habon, 1986] Hornke, L. F. and Habon, M. W. (1986). Rule-based item-bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement*, 10:369–380.

[Kelderman, 1984] Kelderman, H. (1984). Loglinear rasch model tests. *Psychometrika*, 49:223–245.

[Klauer, 1991a] Klauer, K. C. (1991a). Exact and best confidence intervals for the ability parameter of the rasch model. *Psychometrika*, 56:535–547.

[Klauer, 1991b] Klauer, K. C. (1991b). An exact and optimal standardized person test for assessing consistency with the rasch model. *Psychometrika*, 56:213–228.

[Klauer, 1995] Klauer, K. C. (1995). *The Assessment of Person Fit*, pages 97–110. In [Fischer and Molenaar, 1995].

[Klauer and Rettig, 1990] Klauer, K. C. and Rettig, K. (1990). An approximately standardized person test for assessing consistency with a latent trait model. *British Journal of Mathematical and Statistical Psychology*, 43:193–206.

[Klein, 1999] Klein, S. (1999). Vergleich von Item-Antwort-Modellen zur Qualitativen Veränderungsmessung. *Empirische Evaluationsmethoden*, 3:27–46.

[Krause, 1997] Krause, B. (1997). Probleme der Veränderungsmessung im Rahmen der Evaluationsmethodik. *Empirische Evaluationsmethoden*, 1.

[Krzanowski and Marriott, 1995] Krzanowski, W. and Marriott, F. H. C. (1995). *Multivariate Analysis Part 2. Classification, Covariance Structures and Repeated measurements*, volume 2. London.

[Künstner, 2002] Künstner, S. (2002). *Lebensqualität in der Onkologie: Determinanten subjektiver Behandlungstheorien von Krebspatienten*. PhD thesis, Humboldt-Universität zu Berlin, Berlin.

[Lehmann, 1997] Lehmann, E. L. (1997). *Testing Statistical Hypotheses*. Springer, Berlin Heidelberg New York.

[Levine and Drasgow, 1988] Levine, M. F. and Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika*, 53:161–176.

[Levine and Drasgow, 1982] Levine, M. V. and Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology*, 35:42–56.

[Li and Olejnik, 1997] Li, M.-n. F. and Olejnik, S. (1997). The power of rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement*, 21:215–231.

[Lienert, 1973] Lienert, G. A. (1973). *Verteilungsfreie Methoden in der Biostatistik*, volume 1. Hain, Meisenheim.

[Lord, 1980] Lord, F. M. (1980). *Applications of item response theory to practical test problems*. Lawrence Erlbaum, Hillsdale, NJ.

[Lumsden, 1978] Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology*, 31:19–26.

[Maranon et al., 1997] Maranon, P. P., Garcia, M. I. B., and Costas, C. S. L. (1997). Identification of nonuniform differential item functioning: A comparison of mantel-haenszel and item response theory analysis procedure. *Educational and Psychological Measurement*, 57:559–568.

[Masters, 1982] Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47:149–174.

[Meijer, 1994] Meijer, R. R. (1994). The number of guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, 18:311–314.

[Meijer, 1995] Meijer, R. R. (1995). A supplement to the number of guttman errors as a simple and powerful person-fit statistic". *Applied Psychological Measurement*, 19:166.

[Meijer, 1997] Meijer, R. R. (1997). Person-fit and criterion related validity: An extension of the schmitt, cortina and whitney study. *Applied Psychological Measurement*, 21:99–113.

[Meijer, 1998] Meijer, R. R. (1998). Consistency of test behaviour and individual difference in precision of prediction. *Journal of Occupational and Organizational Psychology*, 71:147–160.

[Meijer et al., 1994] Meijer, R. R., Molenaar, I. W., and Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement*, 18:111–120.

[Meijer and Sijtsma, 2001] Meijer, R. R. and Sijtsma, K. (2001). Methodology review: evaluating person fit. *Applied Psychological Measurement*, 25:107–135.

[Meiser, 1996] Meiser, T. (1996). Loglinear rasch models for the analysis of stability and change. *Psychometrika*, 61:629–645.

[Meiser et al., 1998] Meiser, T., Stern, E., and Langeheine, R. (1998). Latent change in discrete data: Unidimensional, multidimensional and mixture distribution rasch models for the analysis of repeated observations. *Methods of Psychological Research Online*, 3:75 –93.

[Mellenbergh, 1982] Mellenbergh, G. J. (1982). Contingency table methods for assessing item bias. *Journal of educational statistics*, 7:105–118.

[Meredith and Millsap, 1992] Meredith, W. and Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, 57:289–311.

[Metzler and Krause, 1997] Metzler, P. and Krause, B. (1997). Methodische Standards bei Studien zur Therapieevaluation. *Methods of Psychological Research Online*, 2:55–67.

[Molenaar and Hoijtink, 1990] Molenaar, I. W. and Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, 55:75–106.

[Nering, 1995] Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement*, 19:121–129.

[Patel et al., 1976] Patel, J. K., Kapadia, C. H., and Owen, D. B. (1976). *Handbook of Statistical Distributions*. Marcel Dekker, New York and Basel.

[Ponocny, 2000] Ponocny, I. (2000). Exact person fit indexes for the rasch model for arbitrary alternatives. *Psychometrika*, 65:29–42.

[Rao, 1973] Rao, C. R. (1973). *Linear statistical inference and its applications*. Wiley, New York.

[Rasch, 1960] Rasch, G. (1960). *Probabilistic Models for some intelligence and attainment tests*. The Danish Institute for Educational research, Kopenhagen.

[Raykov, 1987] Raykov, T. (1987). *Statistische Modelle zur Veränderungsmessung in der Psychologie*. PhD thesis, Humboldt-Universität zu Berlin, Berlin.

[Raykov, 1995] Raykov, T. (1995). Strukturgleichungsmodelle zur Veränderungsmessung in der Psychologie. Berlin.

[Reise, 2000] Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in irt-models. *Multivariate Behavioral Research*, 35:543–568.

[Reise and Widaman, 1999] Reise, S. P. and Widaman, K. F. (1999). Assessing the fit of measurement models at the individual level: a comparison on item response theory and covariance structure approaches. *Psychological Methods*, 4:3–21.

[Rost, 1996] Rost, J. (1996). *Lehrbuch Testtheorie Testkonstruktion*. Huber, Bern.

[Rost, 1999] Rost, J. (1999). Was ist aus dem Rasch-Modell geworden? *Psychologische Rundschau*, 50:140–156.

[Rost and Georg, 1991] Rost, J. and Georg, W. (1991). Alternative Skalierungsmöglichkeiten zur klassischen Testtheorie am Beispiel der Skala 'Jugendzentrismus'. *ZA-Information*, 28:52–75.

[Rost and v. Davier, 1995] Rost, J. and v. Davier, M. (1995). *Mixture Distribution Rasch Models*, pages 257–268. In [Fischer and Molenaar, 1995].

[Roznowski, 1987] Roznowski, M. (1987). Use of tests manifesting sex differences as measures of intelligence: Implications for measurement bias. *Journal of Applied Psychology*, 72:480–483.

[Roznowski and Reith, 1999] Roznowski, M. and Reith, J. M. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement*, 59:248–269.

[SAS Institute Inc., 1999] SAS Institute Inc. (1999). *SAS OnlineDoc®, Version 8,SAS/STAT, Kap. 28, Sect.27.* SAS Institute Inc., Cary,NC.

[Sbandi et al., 1993] Sbandi, P. et al. (1993). *Beschreibung und Bewertung von Evaluationsmethoden im Bereich der Psychotherapie.* Innsbruck.

[Schmitt et al., 1999] Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., and Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement*, 23:41–53.

[Schmitt et al., 1993] Schmitt, N., Cortina, J. M., and Whitney, D. J. (1993). Appropriateness fit and criterion related validity. *Applied Psychological Measurement*, 17:143–150.

[Sedlmeier, 1996] Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. *Methods of Psychological Research Online*, 1:41–63.

[Sijtsma and Meijer, 2001] Sijtsma, K. and Meijer, R. R. (2001). The person-response-function as a tool in person-fit-research. *Psychometrika*, 66:191–208.

[Snijders, 2001] Snijders, T. (2001). Asymptotic null distribution of person-fit statistics with estimated person parameter. *Psychometrika*, 66:331–342.

[Swaminathan and Rogers, 1990] Swaminathan, H. and Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27:361–370.

[Tatsuoka, 1984] Tatsuoka, K. (1984). Caution indices based on item response theory. *Psychometrika*, 49:95–110.

[Trabin and Weiss, 1983] Trabin, T. E. and Weiss, D. J. (1983). *The Person response curve: fit of individuals to item response theory models*, pages 83–108. Academic Press, New York.

[v. Davier and Rost, 1995] v. Davier, M. and Rost, J. (1995). *Polytomous Mixed rasch Models*, pages 371–379. In [Fischer and Molenaar, 1995].

[van Krimpen-Stoop and Meijer, 1999] van Krimpen-Stoop, E. M. L. A. and Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*, 23:327–345.

[Webster and Bereiter, 1963] Webster, H. and Bereiter, C. (1963). *The reliability of changes measured by mental test scores*, pages 39–59. In [Harris, 1963].

[Wilder, 1931] Wilder, J. (1931). Das Anfangswertgesetz. *Zeitschrift für Neurologie*, 137:317–338.

[Witting, 1978] Witting, H. (1978). *Mathematische Statistik: eine Einführung in Theorie und Methoden*. Teubner, Stuttgart.

[Witting, 1985] Witting, H. (1985). *Mathematische Statistik I.* Teubner, Stuttgart.

[Wittmann, 1984] Wittmann, W. (1984). *Die Evaluation von Behandlungs- und Versorgungskonzepten*, pages 87–107. Hogrefe, Göttingen.

[Zwinderman, 1991] Zwinderman, A. H. (1991). A generalized rasch model for manifest predictors. *Psychometrika*, 56:589–600.

Anhang C

EORTC QLQ-C30-Fragebogen (Version 3.0)

EORTC QLQ-C30 (Version 3.0)

Wir sind an einigen Angaben interessiert, die Sie und Ihre Gesundheit betreffen. Bitte beantworten Sie die folgenden Fragen selbst, indem Sie die Zahl ankreuzen, die am besten auf Sie zutrifft. Es gibt keine "richtigen" oder "falschen" Antworten. Ihre Angaben werden streng vertraulich behandelt.

Bitte tragen Sie Ihre Initialen ein:

Ihr Familienname:

Ihr Geburtstag (Tag, Monat, Jahr): . . 1 9

Das heutige Datum (Tag, Monat, Jahr): . . 2 0 0 0

	überhaupt nicht	wenig	mäßig	sehr
1. Bereitet es Ihnen Schwierigkeiten, sich körperlich anzustrengen (z.B. eine schwere Einkaufstasche oder einen Koffer zu tragen?	1	2	3	4
2. Bereitet es Ihnen Schwierigkeiten, einen längeren Spaziergang zu machen?	1	2	3	4
3. Bereitet es Ihnen Schwierigkeiten, eine kurze Strecke außer Haus zu gehen?	1	2	3	4
4. Müssen Sie tagsüber im Bett liegen oder in einem Sessel sitzen?	1	2	3	4
5. Brauchen Sie Hilfe beim Essen, Anziehen, Waschen oder Benutzen der Toilette?	1	2	3	4

Während der letzten Woche:

6. Waren Sie bei Ihrer Arbeit oder bei anderen tagtäglichen Beschäftigungen eingeschränkt?	1	2	3	4
7. Waren Sie bei Ihren Hobbys oder anderen Freizeitbeschäftigungen eingeschränkt?	1	2	3	4
8. Waren Sie kurzatmig?	1	2	3	4
9. Hatten Sie Schmerzen?	1	2	3	4
10. Mußten Sie sich ausruhen?	1	2	3	4
11. Hatten Sie Schlafstörungen?	1	2	3	4
12. Fühlten Sie sich schwach?	1	2	3	4
13. Hatten Sie Appetitmangel?	1	2	3	4
14. War Ihnen übel?	1	2	3	4
Während der letzten Woche:				
15. Haben Sie erbrochen?	1	2	3	4
16. Hatten Sie Verstopfung?	1	2	3	4
17. Hatten Sie Durchfall?	1	2	3	4
18. Waren Sie müde?	1	2	3	4
19. Fühlten Sie sich durch Schmerzen in Ihrem alltäglichen Leben beeinträchtigt?	1	2	3	4
20. Hatten Sie Schwierigkeiten, sich auf etwas zu konzentrieren, z.B. auf das Zeitunglesen oder das Fernsehen?	1	2	3	4
21. Fühlten Sie sich angespannt?	1	2	3	4
22. Haben Sie sich Sorgen gemacht?	1	2	3	4
23. Waren Sie reizbar?	1	2	3	4
24. Fühlten Sie sich niedergeschlagen?	1	2	3	4
25. Hatten Sie Schwierigkeiten, sich an Dinge zu erinnern?	1	2	3	4
26. Hat Ihr körperlicher Zustand oder Ihre medizinische Behandlung Ihr Familienleben beeinträchtigt?	1	2	3	4
27. Hat Ihr körperlicher Zustand oder Ihre medizinische Behandlung Ihr Zusammensein oder Ihre gemeinsamen Unternehmungen mit anderen Menschen beeinträchtigt?	1	2	3	4
28. Hat Ihr körperlicher Zustand oder Ihre medizinische Behandlung für Sie finanzielle Schwierigkeiten mit sich gebracht?	1	2	3	4

Bitte kreuzen Sie bei den folgenden Fragen die Zahl zwischen 1 und 7 an, die am besten auf Sie zutrifft.

29. Wie würden Sie insgesamt Ihren Gesundheitszustand während der letzten Woche einschätzen?

1 2 3 4 5 6 7

sehr schlecht ausgezeichnet

30. Wie würden Sie insgesamt Ihre Lebensqualität während der letzten Woche einschätzen?

1 2 3 4 5 6 7

sehr schlecht ausgezeichnet

Lebenslauf

Name:	Stefan Klein
Geboren am:	12. März 1967 in München
April 1988 - Juni 1995	Studium der Statistik an der Ludwig-Maximilian-Universität München
14. Juni 1995	Abschluss als Diplom-Statistiker
Januar 1996 - Juli 2000	Wissenschaftlicher Mitarbeiter an der Humboldt-Universität zu Berlin, Lehrstuhl Psychologische Methodenlehre, Institut für Psychologie (Prof. Dr. Bodo Krause)
seit August 2000	Beruflich tätig beim Gesamtverband der Deutschen Versicherungswirtschaft (GDV) als Referent in der Abteilung Statistik

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Arbeit selbständig ohne fremde Hilfe verfaßt zu haben und nur die angegebene Literatur und Hilfsmittel verwendet zu haben.

Stefan Klein
23. November 2002